

Politics versus reality

Japan's politicians have to face scientific uncertainty, no matter how uncomfortable it may be. They should mobilize diplomatic means, and not sacrifice scientific integrity, in their fight with North Korea.

The cabinet of Japan's prime minister, Junichiro Koizumi, is "burying its head in its hands" in frustration, in the words of one popular Japanese weekly, over a news article that appeared in *Nature* last month.

At issue is whether Megumi Yokota, a Japanese woman kidnapped by North Korea in 1977 at the age of 13, is still alive. In 2002, North Korea admitted to abducting 13 Japanese nationals, several of them taken from beaches while on dates. Since then, North Korea's half-hearted efforts to account for the victims have caused turmoil in the relationship between the two countries (see *Nature* **433**, 445; 2005).

Claims that most of the victims, including Yokota, have died are unconvincing. North Korea says the remains that it passed to Japan last year are hers. But Japan's tests show that the DNA is someone else's — raising the spectre that the North Korean military is still using her to train spies.

Japan is right to doubt North Korea's every statement. But its interpretation of the DNA tests has crossed the boundary of science's freedom from political interference. *Nature's* interview with the scientist who carried out the tests raised the possibility that the remains were merely contaminated, making the DNA tests inconclusive. This suggestion is uncomfortable for a Japanese government that wants to have North Korea seen as unambiguously fraudulent.

The government has responded sharply to the article. At a press conference, Japan's chief cabinet secretary, Hiroyuki Hosoda, reportedly alleged that *Nature's* article contained "inadequate expressions" and that it misrepresented the scientist's statements. The opinions expressed in the article were "general knowledge" but were not meant to apply to the case at hand, Hosoda said, adding that his statements were checked with the scientist. The scientist himself, meanwhile, is apparently no longer available for interviews.

The inescapable fact is that the bones may have been contaminated. Who knows what they have been through during this hellish

episode? According to North Korea, the body was buried for two years before being dug up and cremated at 1,200 °C, and then kept at the woman's husband's home, before a small sample was passed to Japan. It is also entirely possible that North Korea is lying. But the DNA tests that Japan is counting on won't resolve the issue.

The problem is not in the science but in the fact that the government is meddling in scientific matters at all. Science runs on the premise that experiments, and all the uncertainty involved in them, should be open for scrutiny. Arguments made by other Japanese scientists that the tests should have been carried out by a larger team are convincing. Why did Japan entrust them to one scientist working alone — one who no longer seems to be free to talk about them?

Japan's policy seems a desperate effort to make up for what has been a diplomatic failure — or more precisely, a failure of the security alliance between Japan and the United States. The alliance gives the United States rights to place unpopular bases in Japan in exchange for its role in contributing "to the security of Japan and the maintenance of international peace and security in the Far East".

Could Japan, with US backing, have pulled other levers with North Korea? The answer is not clear, but the question can be put another way. If a totalitarian country had abducted US citizens from a beach and carried them back to teach language to potential spies for 25 years, would George Bush or any other US president be standing there with a bag of ashes haggling over DNA test results?

Part of the burden for Japan's political and diplomatic failure is being shifted to a scientist for doing his job — deriving conclusions from experiments and presenting reasonable doubts about them. But the friction between North Korea and Japan will not be decided by a DNA test. Likewise, the interpretation of DNA test results cannot be decided by the government of either country. Dealing with North Korea is no fun, but it doesn't justify breaking the rules of separation between science and politics. ■

Nature respects preprint servers

... despite false rumours to the contrary.

The surest way of hiding the truth about a journal's policies, it seems, is to publish it in the Guide to Authors. At *Nature*, this rich treasure trove of essential information is lovingly cared for by staff. Successive improvements in our presentation of research papers are reflected in changes to stylistic guidelines. Editors labour mightily over policy development, and craft their statements with all the precision of a Swiss watchmaker.

Alas, this is all to no avail. Too often, authors submit papers that fail to conform to our guidelines; outraged readers wishing to complain about content fail to follow the procedures; and journalists spread false representations of our media policies.

But perhaps the most malignant rumours are those spread by scientists themselves — in all good faith but erroneously. One such is about prepublication. *Nature*, it is said, does not allow authors to distribute preprints of their work or place drafts of their papers on

preprint servers, or even discuss their submitted work at conferences.

This is false.

Nature never wishes to stand in the way of communication between researchers. We seek rather to add value for authors and the community at large in our peer review, selection and editing.

Communication between researchers includes not only conferences but also preprint servers. The ArXiv preprint server is the medium of choice for (mainly) physicists and astronomers who wish to share drafts of their papers with their colleagues, and with anyone else with sufficient time and knowledge to navigate it.

So please let's put a myth about this journal to rest. As first stated in an editorial in 1997, and since then in our Guide to Authors, if scientists wish to display drafts of their research papers on an established preprint server before or during submission to *Nature* or any *Nature* journal, that's fine by us. ■





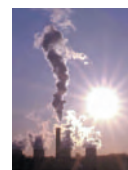
Australian rules
A rugby team
tailors training
with genetic tests
p260



Space suit
Bush plumps for
Michael Griffin
to head NASA
p261



Up on the farm
Greenhouse-gas
emissions may be
underestimated
p262



Quit smoking
Britain settles
for a lower
carbon allowance
p264

Indian regulations fail to monitor growing stem-cell use in clinics

K. S. Jayaraman, New Delhi

Two Indian government departments that have fought for years over who should control stem-cell research are finally negotiating a single set of guidelines for clinical practice.

The move has been driven by the mushrooming of clinics, both public and private, that claim to use stem cells to treat a wide range of diseases. The cells come from bone marrow, umbilical-cord blood and even human embryonic sources. Yet there is no supervision of procedures for preparing stem cells or clinical follow-up, says Vasantha Muthuswami, head of basic medical sciences at the Indian Council of Medical Research (ICMR) in New Delhi.

"We want to promote stem-cell technology but not in this scandalous way," Muthuswami says. Basic research to prove the efficacy and safety of procedures has not been completed anywhere in the world, she points out.

In 2002, the ICMR, which belongs to the health ministry, announced a policy that permitted therapeutic cloning and encouraged stem-cell research. But the previous year the Department of Biotechnology (DBT), which belongs to the science ministry, had also issued guidelines, and some clinics had exploited these to begin clinical treatments.

The ICMR has received numerous applications for funding in recent months, some of which are startlingly naive, says Muthuswami. But it has also received notification of ongoing clinical research, seemingly approved by the DBT and local ethical committees, and she suspects that at least one of these involves stem cells derived from human embryos. Four institutions have informed the ICMR that they are planning to start transplants of human embryonic stem cells. Private blood banks are being established to supply therapies using cord-blood stem cells, she adds. And many private clinics that are offering stem-cell therapy have never contacted the ICMR.

No one has a clear idea of what clinical studies are being carried out where, and how they are being evaluated, admit the ICMR and the DBT. Now the head of the DBT, Mahiraj Bhan, who took office last year, has

Moving too fast: the All India Institute of Medical Sciences is already using stem cells to treat stroke.

asked for a meeting with the ICMR to agree common guidelines to ensure clear supervision and appropriate controls. And on 5 March at a meeting in Hyderabad, the head of the DBT's stem-cell task force, Dorairajan Balasubramanian, announced plans to conduct a series of workshops to familiarize doctors with rules on good clinical practice.

Hard sell

Hyderabad seems to be a centre for stem-cell medicine — the privately owned L. V. Prasad Eye Institute in Hyderabad pioneered techniques as early as 2001. The institute says it has used transplanted stem cells to treat more than 240 patients with damaged corneas. Two other private hospitals in Hyderabad are using stem-cell therapy to treat damaged heart muscle, and plans are in place in other hospitals and clinics in the city to use stem cells to regenerate the liver in cirrhosis sufferers and the pancreas in diabetics.

One of the biggest shocks has been last month's announcement in the *Times of India* that a top heart surgeon at New Delhi's All India Institute of Medical Sciences (AIIMS) has used stem cells derived from bone marrow to treat 35 patients during bypass

surgery. The surgeon, Panangipalli Venugopal, said in the newspaper report that he and his colleagues had also administered stem cells to patients with cerebral palsy, muscular dystrophy and stroke. Venugopal, who has not published his work in a peer-reviewed scientific journal, was not available to comment.

"We are only a block away from AIIMS and we did not know this was happening there," says Muthuswami. "If the nation's premier medical institute did not ask our permission for such therapy, how can we blame private clinics for what they do?"

The quality of cells being used in therapy is of major concern, as is the failure of clinicians to understand stem-cell biology. "A lot of basic research is needed and the safety and efficacy of therapy must be experimentally proven in animals," Polani Seshagiri, a stem-cell biologist at the Bangalore-based Indian Institute of Science, told *Nature*. "The cells we put in the body must be of clinical grade, and how many clinicians ensure this?"

Muthuswami says that the planned guidelines will not stop research but will ensure that there are sufficient checks and controls on clinical practice. "It will no longer be a free-for-all."

Rugby team converts to give gene tests a try

Carina Dennis, Sydney

An Australian rugby league team claims it has gained a competitive edge over its rivals by using genetic tests to tailor its players' training programmes. This move marks the beginning of more widespread use of genetics in sport, according to geneticists and legal experts.

"Once one club has it, there will be pressure on other teams to do everything they can or risk falling behind the technology," says David Weisbrot, president of the Australian Law Reform Commission, which makes legislative and policy recommendations regarding genetic information to the nation's government.

The Sea Eagles, a professional rugby team in Manly, a seaside suburb of Sydney, has DNA-tested 18 of its 24 players for 11 exercise-related genes. Some of the genes tested are associated with the efficiency of oxygen use; others are involved in muscle development and clearing lactate, the chemical that causes muscle stiffness.

Each player's training programme is being redesigned according to whether his individual genetic profile indicates, for example, that his main strength is endurance or speed. "We don't want a player running 100 kilometres a week if his genes say he's more suited to 50 kilometres a week and

Hard to tackle: experts fear there will be a scrum over gene technology.

doing more weights," says Steve Dank, the team's physiologist behind the programme.

But applying genetics to sports training is premature, say some experts. "We don't know enough yet about these genes," says Ron Trent, a molecular geneticist at the University of Sydney.

Trent also argues that genetic tests are likely to be "more meaningful at the elite level, where it could be the difference between a gold and silver Olympic medal". But he agrees that they could help to identify an athlete's potential.

He is working with the Australian Institute of Sport (AIS), the country's premier sports training and research facility, to develop genetic profiles that identify the sport best suited to an athlete. Genetic indicators of speed, power and endurance can inform decisions about whether an athlete is better suited to swimming or rowing, for instance.

Trent's team is also investigating genetic profiles to position rugby union players on the field — for example, whether they are better suited to being nimble backs or beefy forwards.

There could soon be pressure on sports clubs, for financial and insurance reasons, to extend their testing to screen for genetic predisposition to certain injuries,

Weisbrot says. "I also have concerns about privacy and genetic discrimination," he adds.

DNA samples are destroyed immediately after testing and the team has agreements in place to stop the information being obtained by other parties, according to Dank.

Dank believes that genetic screening will become routine for sports clubs — "I've had a number of clubs contact me," he says. But he doubts that they will ever be critical factors in selecting players. "Gene tests don't measure the passion that makes players great," he says. ■

Japan plans blood-donor restrictions to combat vCJD

Ichiko Fuyuno, Tokyo

Japan is considering a tough new rule to prevent the spread of new-variant Creutzfeldt-Jakob disease (vCJD), the human equivalent of mad cow disease.

If the rule passes when an advisory committee addresses it again in a few months' time, it would ban anyone who spent even a single day in Britain or France between 1980 and 1996 from giving blood. This would further tighten current restrictions, which bar people who have spent more than one month in those countries, and is expected to hugely reduce the number of eligible donors.

Other countries have taken similar, if less cautious, measures. The United States, for example, does not allow people who have lived in Britain for several months to give blood. But Japan's new rule, proposed by an advisory committee to the ministry of health on 7 March, is the most stringent yet. "We are taking the safest possible measure until

the risk is further determined," says Daisaku Sato, deputy director of the ministry's blood and blood-products division.

The proposed change reflects fears that arose after ministry officials announced the first confirmed case of vCJD in Japan in early February: the man was thought to have been infected during a stay of less than a month in Britain around 1990. He began to show symptoms in 2001, and died in December last year.

About 160 people have died from vCJD around the world, mostly in Britain. Until now, most fatalities have occurred among people who had lived in the country for at least a few years. The disease is thought to be caused mainly by eating meat infected with rogue proteins called prions, but there have been at least two cases of probable infection from blood transfusions.

Government officials and scientists are keen to determine how many people are carrying the disease without symptoms, and

what the chances are that such people may transmit the disease through transfusions or surgery. In Britain, steps have been taken to reduce the risk — such as removing white blood cells from transfusion products. British blood banks are also considering filtering blood for prions, or restricting donors to low-risk age groups. But these options are expensive and severely limit blood supplies.

The Japanese Red Cross Society estimates that the new rule would cost it hundreds of thousands of potential donors. "The damage would be significant," says a spokesman.

And Masahito Yamada, a prion-disease specialist at Kanazawa University, says that the ministry's efforts may have a limited effect as Japan relies on imports for many blood products. This poses concerns that contaminated blood could still enter Japan. "It is really important to establish a safe global distribution system for blood products," Yamada says. ■

Bush settles on technical innovator to head up NASA

Tony Reichhardt

Michael Griffin, an engineer and physicist who builds planetary spacecraft, is the White House's popular selection for NASA's 11th administrator.

If the appointment is confirmed by the Senate, it will be Griffin's second stint with the US space agency — he served as NASA's chief engineer and the head of its exploration office in the early 1990s. After that he worked for aerospace companies and for In-Q-Tel, a Virginia-based company that develops technology for the US Central Intelligence Agency. Since last April Griffin has headed the space department at the Johns Hopkins University Applied Physics Laboratory in Laurel, Maryland. The lab has become a main supplier of spacecraft for NASA planetary missions, such as the MESSENGER mission to Mercury and the New Horizons project to visit Pluto.

In the 1980s, as deputy for technology for the Strategic Defense Initiative, Griffin pioneered the concept of using small, smart spacecraft instead of big, expensive ones. This culminated in 1994's Pentagon-led lunar orbiter, Clementine, and influenced NASA's switch to 'better, cheaper, faster' missions. As the agency's associate administrator for exploration, he brought the same philosophy to the short-lived Space Exploration Initiative, a proposal by the first President George Bush to send astronauts to the Moon and Mars. Griffin turned to centres outside NASA for lunar exploration proposals in the hope of lowering mission costs.

Unlike his predecessor Sean O'Keefe, whose background is in accounting and government, Griffin is a space technologist. His roots lie in Maryland, the home state of Senator Barbara Mikulski, a Democrat and strong advocate of space science. Mikulski battled with O'Keefe over his decision to cancel servicing of the Hubble Space Telescope, and the question of whether to reverse that decision is one of many facing Griffin.

He inherits an agency that many critics say is too stodgy, and is not configured to carry out the current president's Moon-Mars initiative — just last week NASA announced plans to retire up to 15% of its workforce. Among the dilemmas facing the agency are how and when to retire the space shuttle, and what to do about the still uncompleted International Space Station.

Last year, Griffin co-chaired a study on human space exploration for the Planetary



Michael Griffin: scientists praise the decision to appoint this pioneer of small, smart missions.



Society, which is based in Pasadena, California. The study recommended that NASA consider ending shuttle flights before the full station is built to save money for the exploration programme. In congressional testimony, Griffin said: "It is beyond reason to believe that [the space station] can help to fulfill any

objective, or set of objectives, for space exploration that would be worth the \$60 billion remaining to be invested in the program."

Many in the field of space exploration welcome Griffin's technical know-how and his willingness to question old ways of doing business. Wesley Huntress, a former NASA head of space science who directs the Geophysical Laboratory at the Carnegie Institution of Washington, calls him "an honest person who says what he thinks".

News of Griffin's nomination won immediate praise from scientists and politicians, Democrats and Republicans alike. Alan Stern is a planetary scientist with the Southwest Research Institute in Boulder, Colorado, whose New Horizons mission is being built by Griffin's Applied Physics Laboratory. He says, "I really think he will be the best administrator that NASA has had in decades. I expect he will make NASA shine."

Global bomb-test monitor could give tsunami warnings

Declan Butler, Paris

Real-time data from seismic stations being built to detect nuclear tests may be made available to a proposed tsunami warning system.

The International Monitoring System (IMS) is currently used by a Vienna-based section of the Comprehensive Nuclear Test Ban Treaty Organization (CTBTO) to check for unauthorized nuclear-weapons tests.

Lassina Zerbo, director of the organization's International Data Centre, announced the possible new use of its data at an international meeting at UNESCO in Paris from 3 to 8 March. The gathering was held to discuss technical aspects of the creation of a tsunami warning system for the Indian Ocean.

The IMS is potentially extremely useful. Its planned global network of 321 monitoring stations — of which some 130 are already in operation — will use seismic, hydroacoustic (underwater sound) and infrasound (sound frequency below that detectable by the ear) data to measure tremors. Data can also be relayed in real time around the globe by the CTBTO's satellite infrastructure.

But the move is highly sensitive politically. The CTBTO will make data available to UNESCO's Intergovernmental Oceanographic Commission (IOC), which will not pass the information to any other body or individual.

Initially the IOC will test the types of data that could be useful, and look at how these could be integrated into a warning system.

"We'll see in the next three months whether the data can be used for disaster warning systems," says the IOC's executive secretary, Patricio Bernal.

The Paris meeting also saw the agreement of a timetable for creating a system of tide gauges and seafloor pressure monitors for the Indian Ocean. By the end of this year, six new monitoring stations will be installed off the coasts of Malaysia, Thailand and Indonesia, and 15 others in the region will be upgraded to monitor sea levels.

By the end of this month, the United States and Japan will provide alerts on all seismic activity in the Indian Ocean region to round-the-clock contact points in the surrounding countries. But discussions on the final form and cost of the Indian Ocean system could take until the end of 2006, says Bernal.

Fallout of fertilizers set too low, studies warn

Jim Giles, London

The internationally agreed technique used to estimate levels of a potent greenhouse gas may be greatly underestimating emissions, researchers have told *Nature*.

Emissions of nitrous oxide resulting from fertilizer use in Britain may be up to 50% higher than estimated by current guidelines, according to recent studies. The guidelines may also be hampering attempts to limit emissions, as they fail to reward countries for switching to fertilizers that release less nitrous oxide.

Uncertainties surrounding nitrous oxide levels are greater than for any other greenhouse gas covered by the Kyoto Protocol on climate change. In terms of total emissions, the gas has the greatest global-warming potential after carbon dioxide, roughly on a

par with methane. But calculating releases from agricultural soils — the single biggest nitrous oxide source — is difficult because of the many factors that affect how microbes break down nitrogen-containing fertilizers.

Most governments instead stick to guidelines produced in 1996 by the Intergovernmental Panel on Climate Change (IPCC), which estimate that 1.25% of nitrogen applied as fertilizer is emitted as nitrous oxide. Use of a single rough estimate is judged to be the only solution that will work internationally, given that local variables such as soil type and humidity level significantly affect nitrous oxide production.

But the IPCC encourages governments to use more sophisticated techniques if appropriate peer-reviewed literature becomes available. And there are now many such data.

Researchers in Britain in particular have conducted some of the most detailed studies on nitrogen use, and these put Britain's total nitrous oxide emissions at up to twice the IPCC estimate. The researchers and their colleagues in Germany — where studies suggest that emissions from some agricultural areas may also be twice as high as IPCC estimates — have now told their governments that the IPCC methodology underestimates emissions and needs to be updated.

In one of the most recent studies, to be published later this year, Keith Smith, an environmental scientist at the University of Edinburgh, used a combination of empirical data and modelling to find that annual nitrous oxide emissions from fertilizer use in Scotland have reached 17,000 tonnes, more than half as much again as the 1.25% estimate. These results tally with nationwide models developed in the past five years by Ute Skiba, a soil microbiologist at the Centre for Ecology and Hydrology in Penicuik, UK.

The British and German environment ministries are now considering updating the methodologies they use, and Germany may even implement new guidelines next year. The situation is not the same in all countries, however. Studies in France and Hungary, for example, suggest that 1.25% is a good estimate.

UK officials add that because the IPCC system is based on total fertilizer use, it deters countries from investigating a switch to varieties of nitrogen fertilizers that produce less nitrous oxide. In the meantime, they add, the ministry is working with farmers to promote more efficient use of fertilizer.

Additional reporting by Tamara Grüner.

Farm produce: emissions of nitrous oxide from fertilizers may be 50% higher than current estimates.

Gene tests served up to tell fine foods from fakes

Rachael Williams, Tokyo

Japan is set to crack down on imitations of its local delicacies — from seaweed to crab — by using genetic testing to check the origins of foodstuffs.

Many Japanese products use their place of origin as a sign of quality, in the same way that real champagne must come from the region in France of the same name. But imitations often end up on the market.

Government officials in Hokkaido, Japan's largest food-supplying region, say that tofu, soba noodles, beans and other products are sometimes sold under a fake Hokkaido label. But in April, the local government will start to protect the region's economic interests by implementing spot checks of products in supermarkets across the country.

The idea of genetic testing of food is not new. The Kyoto-based company Takara Shuzo, for example, has developed a genetic test that can distinguish famous Koshihikari rice — from the Niigata region — from cheaper varieties. It sells kits to retailers so they can check the authenticity of their products.

A similar test has been developed by researchers at the National Institute of Vegetable and Tea Science in Mie Prefecture to allow them to distinguish genuine Tochiotome strawberries from Korean fruit that illicitly bears the Tochiotome label.

And Japan's customs agents last year started using genetic tests on imports of tatami mats — a traditional Japanese floor covering. This should ensure that mats described as Hinomidori are made from the

high-quality reeds developed by researchers at the Kumamoto Prefecture Agricultural Research Center.

The challenge in store for Hokkaido will be to identify genetic markers unique to all the products they wish to test. Finding markers for Hokkaido's crab, for example, will be quite a challenge, says Hideyuki Imai, who studies crab genetics at the University of the Ryukyus in Okinawa. If crabs from different regions interbreed with each other, says Imai, it may prove to be impossible.

All this rigorous genetic testing may force up the price of these items, notes Yoshito Tsuya, an agricultural economist at Utsunomiya University, just north of Tokyo. But he thinks that customers will be prepared to pay more to make sure they get the real thing.

Plans for research watchdog praised, but it may lack teeth

Jim Giles, London

British biomedical science is to get its first watchdog. But even as details of the plan are being finalized, its architects have generated controversy by saying that they would consider accepting funding from the pharmaceutical industry.

Critics also fear that the proposed UK Panel for Health and Biomedical Research Integrity — which will operate under the umbrella of Universities UK (UUK) — may lack the teeth to police research properly.

Plans to establish the panel should be completed next month at a board meeting of the UUK, which represents most of Britain's higher-education institutes.

The idea of the panel has been spearheaded by Michael Farthing, principal of St George's Hospital Medical School in London. It has the approval of the government, and the Department of Health is currently helping to select the first panel members. The body will initially cover all biomedical research in universities and the National Health Service.

The need for an independent body to oversee investigations of misconduct, such as plagiarism and fraud, has been made clear by several science organizations in recent years.

But although the panel is likely to act as a safe haven for whistle-blowers, it will not itself investigate misconduct allegations. John Pritchard, a senior policy adviser at Sheffield Hallam University who is working on the plans on behalf of UUK, says the panel will instead advise institutions on how to run inquiries and nominate participants.

"I'd like to see a rigorous system of investigation and enforcement," says Brian Gennery, president of the Faculty of Pharmaceutical Medicine at the Royal Colleges of Physicians, one of the first UK organizations to promote the need for a research-integrity watchdog. "Then I would have much more confidence in this new body."

Some organizations established elsewhere to oversee research integrity, such as those in Denmark and the United States, have the power to run their own inquiries. Pritchard declined to comment on this issue until details of the plan have been approved by UUK, saying only that the whistle-blower facility alone would have a positive effect on subsequent investigations by employers.

The panel will also develop a non-mandatory code of conduct for biomedical researchers and raise awareness of misconduct through local events.

If UUK approval is secured next month, Pritchard expects panel members to be nominated in May and to start work by October.

The panel would initially be established for a three-year period, pending a review of its performance.

Pritchard discussed the plans on 11 March at a London meeting of the Committee on Publication Ethics, an association of journal editors. When Pritchard mentioned that UUK was considering accepting funding from a range of sources — including the pharmaceutical industry — some delegates responded angrily.



Bright idea: Universities UK will discuss who funds a new body to oversee biomedical research.

"How will pharmaceutical sponsorship square with public confidence in this body?" asks Iona Heath, a London-based general practitioner and chairwoman of the *British Medical Journal's* ethics committee. She points out that the behaviour of the industry is currently under investigation by parliament following claims that some firms failed to publish sensitive research (see *Nature* 429, 793; 2004). "This is not the right time to jump into bed with them," says Heath.

Pritchard says no decision on funding has been taken, but adds that the pharmaceutical industry is a major funder of biomedical research and it would be a mistake not to engage with them. "They're a stakeholder," he says. "It will be a missed opportunity if their responsibility is put aside."

Row flares over attempt to dilute radioactive waste

Emma Marris, Washington

Independent experts have finally been given the chance to comment on highly contentious plans for disposing of almost 100 million gallons of radioactive waste.

The waste was generated during the cold war when spent fuel rods from nuclear reactors were converted to weapons-grade plutonium. It is now being kept in massive tanks at Department of Energy sites in Idaho, South Carolina and Washington state.

The problem of its disposal has been a bone of bitter political contention for decades. Particularly rancorous is the issue of whether the energy department should have the power to reclassify certain waste so that it can deal with it more flexibly and cheaply. This power was successfully challenged by an environmental group, the Natural Resources Defense Council, in 2003.

The department regained that power this year in Idaho and South Carolina, when Congress's Armed Services Committee inserted a provision into the National Defense Authorization Act to make it legal. But under pressure, the committee conceded that 21 independent experts at the National Academy of Sciences should be allowed to assess the technical risks in the agency's proposed disposal plan.

The plan calls for various treatments for the contents of the storage tanks. Liquids that are less radioactive would be dealt with on-site. The radioactive salt cake would be extracted, processed and shipped to a long-term repository. Concrete would then be added to the highly radioactive sludge that has settled at the bottom of the tanks to dilute the radioactivity.

It is this last step, which critics call 'disposal by dilution', that the experts, who met on 7 March, are looking at. They will assess the long-term safety of the technique and will examine alternatives.

Environmentalists are angry that the energy department has regained freedom from rigorous oversight in nuclear-waste disposal. They hope that the panel will expose what they see as an inadequate clean-up and a lack of credibility.

Energy department spokesman Joe Davis denies the lack of credibility. "Many agree that this is the way to proceed," he says, adding that the agency reviews reports from the National Academies and "takes their opinions and thoughts into account as we go about our work".

UN compromise ends human cloning debate with 'non-binding' ban

Washington The United Nations has ended an acrimonious stand-off over cloning by agreeing to a non-binding declaration. The result lets both sides of the debate claim victory, but will probably not have a great impact on research.

On 8 March, the UN General Assembly voted by 84 to 34, with 37 abstentions, to pass a declaration calling on all its member states to ban all forms of cloning, including both reproductive and therapeutic forms.

The vote caps more than three years of debate. In 2001, France and Germany initiated an effort to ban reproductive cloning, to which the United States responded by lobbying for a ban on research work too. The ensuing debate deadlocked the UN, ultimately leading to this non-binding declaration as a compromise.

Opponents of therapeutic cloning can claim a moral victory, observers say, although nations that want to carry out research as before will be free to do so. The fact that the UN did not take a legal stand on reproductive cloning dismays many on both sides of the issue.

Britain retreats in dispute over carbon emissions

London Britain has ended a stalemate with the European Commission over limits on UK carbon emissions.

The UK government originally asked the commission for allowances of 736 million tonnes of carbon dioxide per year — an amount that was agreed last July. Britain later decided it needed more, thanks to rising power-generation requirements, and

asked for 756 million tonnes. But the commission was not convinced.

Britain has withdrawn its request and settled on the lower figure, but is planning a legal appeal. The climbdown means that Britain can now trade greenhouse-gas allowances as part of the European Emissions Trading Scheme, which began on 1 January (see *Nature* 432, 268–270; 2004).

Blair panel calls for doubling of African aid

London World leaders are to debate plans to invest US\$8 billion over ten years in African science and technology. The sum is part of a proposed package to boost the continent's development. The package also suggests doubling aid levels over three to five years, eventually spending an additional \$50 billion a year.

The proposal will be discussed by the heads of the eight leading industrialized nations (G8) in Scotland this July. Similarly ambitious plans have been mooted before, but this scheme is attracting interest because it comes from the potentially influential Commission for Africa. This 17-member panel, consisting mainly of African representatives, was set up last year by Tony Blair, Britain's prime minister and current G8 president, to guide this summer's talks.

Climbing high? Britain wants to increase its emission allowances for power stations.

Duo leave government jobs after conflict hearings

Washington Two scientists who testified before Congress about possible conflicts of interest in their work have left their government jobs for roles in academia.

Lance Liotta and Emanuel Petricoin were employed by the US National Institutes of Health (NIH) and the Food and Drug Administration, respectively, when they were asked to testify before a congressional committee last May about their consultancy work. The pair have also had to defend themselves against claims that a cancer test based on their research is not as effective as they claimed (see *Nature* 429, 496; 2004).

On 9 March, George Mason University in Manassas, Virginia, announced that Liotta and Petricoin have both accepted positions there, starting this spring.

Observers say the departures are no surprise after last year's rough ride. Some predict that many more scientists will leave the NIH in the wake of tough new limits on consulting work (see *Nature* 433, 557; 2005). And on 9 March, a group of NIH scientists calling themselves the Assembly of Scientists made a counter-proposal for softer limits. Without such a change, said one observer who declined to be named, "all the best and the brightest will leave".

Aquarium's prize exhibit shows killer touch

San Diego An experiment at the Monterey Bay Aquarium aimed at keeping a great white shark in captivity is proving to be a success — but at the expense of the soupfin sharks.

The great white, introduced into a 4-million-litre tank at the California aquarium in September (see *Nature* 431, 233; 2004), has in recent weeks killed two soupfins kept in the tank.



Aquarium officials say the soupfins didn't steer sufficiently clear of the great white (pictured below), provoking it to attack. Sean Van Sommeran, who runs the Pelagic Shark Research Foundation in nearby Santa Cruz, contends that the turf war is a sign of stress on the great white, which is known to roam 100 kilometres a day in the ocean. But aquarium officials say that

the shark, kept alive in captivity longer than any other, is thriving. If problems emerge, they say, it will be released.

The great white has boosted the aquarium's once flagging attendance by some 30%.

NSF to deter applicants for more grant success

Washington The National Science Foundation (NSF) is taking steps to cut paperwork and delays in its grant procedures — by reducing the number of applications it receives.

Requests for proposals will soon be directed and technically specific, says the NSF's director Arden Bement. "Hopefully this will reduce the proposal volume," he told the US House science committee in a hearing

on 9 March. The change is needed, Bement says, because the NSF is too understaffed to handle the roughly 44,000 requests it now receives each year.

The move should have the fortunate side-effect of raising the success rate for grant proposals from its current level of about 20% — something on which Bement previously promised to focus. This cannot be done by handing out more grants because the NSF's budget of \$5.6 billion in 2004 was cut this year by 2%.

MONTEREY BAY AQUARIUM



The X factor

The sequence of the 'feminine' X chromosome is a prime hunting ground for geneticists interested in the evolution of the cognitive and cultural sophistication that defines the human species. Erika Check reports.

This January, Harvard University president Larry Summers incited a near riot by suggesting that men might be better than women at science. The resulting pandemonium has revealed few genuine insights into male and female mental abilities — although it has shown that old prejudices linger on campus, and beyond.

In contrast, biology presents a challenge to those who still believe women are better off at home than in the hallowed halls of universities. As geneticists search for the roots of humanity's unique mental abilities, they are beginning to pay close attention to the 'feminine' X chromosome. Women have two copies of this chromosome, whereas men have only one. And the complete sequence of the X chromosome, published in *Nature* this week¹ (see also News and Views, page 279), confirms that an unusually large number of its genes code for proteins important to brain function.

Why this should be the case is sparking debate among evolutionary biologists. And some are even suggesting that the X chromosome will tell us why we are different from our closest relatives — why we can write poetry and design nuclear weapons, but chimpanzees can't. In a sense, they argue, the feminine chromosome could hold the secrets of humanity. "We used to think that the X was boring," says Jenny Graves, an evolutionary geneticist at the

Australian National University in Canberra. "Now we're seeing just how wrong we were."

Today's understanding of the X chromosome helps to explain a puzzling observation from the late nineteenth century, when doctors combing through data from the 1890 US Census noticed that more boys than girls were mentally disabled². We now know that this reflects a preponderance of genes for brain function on the X chromosome. A woman uses only one of her two X chromosomes in each cell, so if one of her X chromosomes has a defective gene, only some of her cells will suffer. But men have only one X, so any defective brain genes from that chromosome are invariably expressed.

All in the mind

Many different types of mental retardation have since been linked to defects in genes on the X chromosome — far more than can be explained by their chance distribution throughout the genome. According to one analysis, there are 221 known human genetic defects that can cause mental impairment, some 10% of which reside on the X chromosome, even though it carries less than 4% of known human genes³.

Detailed studies have also shown that the specific genes linked to mental disabilities

play crucial roles in normal brain function⁴. For instance, more than a decade ago, an international team reported the discovery of the gene that causes fragile X syndrome⁵, a disorder that leads to a range of problems including mental disability. Scientists now know that the defective gene, called *FMRI*, normally makes a protein that is involved in shuttling the genetic messages that enable nerve cells to send signals through the brain⁶. Another gene, known as *MECP2*, leads to a whole range of mental disorders when mutated; its protein seems to be involved in the silencing of other genes required for normal learning, memory and the growth of brain cells⁴.

It's still a giant step from understanding defects in single genes to proving that the X chromosome allows us to write novels and solve calculus equations. But there is some indirect

evidence that genes on the X chromosomes are involved in higher cognitive functions. One hint comes from a study of 4,000 sets of British identical twins. Each female twin inherits two X chromosomes, one from her mother and one from her father, but each individual twin randomly inactivates one of her two X chromosomes. So identical twin sisters can express different X-chromosome genes. In contrast, male identical twins

"If higher cognitive abilities were a critical step in our evolution, it makes sense that you'd find those functions on the X chromosome."

— Hunt Willard

inherit only one X chromosome, from their mothers, and so must activate the same X-linked genes. In the British study, researchers led by Ian Craig of King's College London found that in some traits linked to intelligence, such as verbal skills and good social behaviour, male twins were more alike than female twins⁷.

But why should the X chromosome have emerged as a hotspot for genes influencing our cognitive abilities? Evolutionary geneticists believe that the two mammalian sex chromosomes, X and Y, were once identical. As mammals began to diverge from their reptilian ancestors, some 300 million years ago, the proto-X and proto-Y chromosomes took on the role of determining an individual's sex. Both initially started accumulating genes from elsewhere in the genome, but over time the two chromosomes began to grow apart; the Y started to shrink and lost many of its genes.

Solid foundation

Eventually, the sex chromosomes diverged to the point where they were no longer able to exchange genetic material during the cell divisions that give rise to sperm and eggs, as do the members of every other pair of chromosomes. This has left the X chromosome as one of the most stable in the mammalian genome — which paradoxically may have exposed its genes to more intense pressure to evolve.

The X chromosome gets a chance to shine, or to fail miserably, each time it passes through the male line. Because a male carries only one copy, any new mutations are revealed in all their glory. And because successful males have the potential to sire very large numbers of children with multiple partners, mutations on the X chromosome that are advantageous to both sexes can spread rapidly through a population.

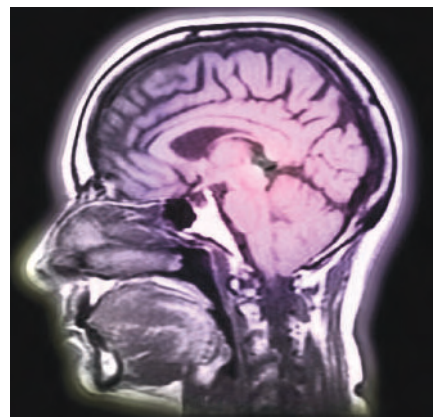
In our own species, where intelligence and social skills are thought to be central to success, genes on the X chromosome seem to have evolved rapidly to provide us with the necessary brain power. "If higher cognitive abilities were a critical step in our own evolution, it makes sense that you might find those functions on the X chromosome," says Hunt Willard, a human geneticist and director of the Institute for Genome Sciences and Policy at Duke University in Durham, North Carolina.

Provocatively, researchers led by Horst Hameister at the University of Ulm in Germany speculate that this process was driven by sexual selection³. Early in human evolution, they suggest, females developed a preference for intelligent males. According to their theory, the genes for super-intelligence and for the preference of intelligent males were closely linked, and so were inherited together. And because superior intelligence also aided survival, the process wasn't kept in

Double vision: by looking for differences in the behaviour of twin girls, geneticists are trying to find out how X chromosomes (top) influence mental ability.

check by natural selection — unlike other sexually selected characteristics such as the peacock's tail, which makes its bearers more vulnerable to predators.

Many of the genes on the X chromosome associated with human brain function seem to have distant relatives with different functions in other vertebrates, such as chickens and fish⁸. So in boosting our cognitive abilities, the X chromosome seems to have co-



Head start: our large brains may have evolved thanks to the action of the X chromosome.

opted a diverse range of existing genes, rather than evolving a new set of genetic sequences for the purpose. "These old genes are getting new use," says Hameister.

In some instances, geneticists have pinpointed genes on the X chromosome that still seem to be in the process of adopting new roles in the brain. For instance, a gene called *JARID1C* seems to be evolving from a similar gene called *JARID1D*, which is found on the Y chromosome. If men inherit a damaged version of the *JARID1C* gene on their single X chromosome, they develop mental disabilities. The fact that the healthy Y chromosome version cannot compensate for its defective cousin hints that *JARID1C* is becoming more crucial to the brain as it evolves⁹.

Personality profiles

Geneticists are now gearing up to go after other X-linked genes that may help explain what makes us human. In London, Craig's team plans to identify twins who score high or low on certain 'people skills', such as sharing their toys and volunteering help to others. The researchers will then use gene chips to scan the twins' DNA, looking for particular genetic variations that correlate with these traits. Once they find a region of DNA that seems to link to a particular trait, the group will look at the detailed sequence of individual chromosomes to try to pin down the exact gene involved. The X chromosome data, with its wealth of information about human brain genes, is likely to feature prominently in this endeavour.

Other researchers plan to continue the quest for genes involved in X-linked brain disorders. Every two years, for instance, scientists meet as part of a European consortium that catalogues genes involved in such conditions. Researchers believe that the information so far gleaned about human brain function from these studies barely scratches the surface.

If we want to understand the cognitive 'X factor' that separates us from the rest of the animal kingdom, then it seems that the X chromosome is the place to start looking. In the meantime, Summers and his acolytes can chew on this thought: even if there's any truth in the idea that men are more suited to a career in science than women, they just might owe this mental predisposition to the 'girly' chromosome. ■

Erika Check is Nature's Washington biomedical correspondent.

1. Ross, M. T. *et al. Nature* **434**, 325–337 (2005).
2. Johnson, G. E. *J. Psycho-Aesthetics* **2**, 26–32 (1897).
3. Zechner, U. *et al. Trends Genet.* **17**, 697–701 (2001).
4. Ropers, H.-H. & Hamel, B. C. *J. Nature Rev. Genet.* **6**, 46–57 (2005).
5. Verkerk, A. J. *et al. Cell* **65**, 905–914 (1991).
6. Oostra, B. A. & Chiurazzi, P. *Clin. Genet.* **60**, 399–408. (2001).
7. Loat, C. S., Asbury, K., Galsworthy, M. J., Plomin, R. & Craig, I. W. *Twin Res.* **7**, 54–61 (2004).
8. Kohn, M., Kehrer-Sawatzki, H., Vogel, W., Graves, J. A. M. & Hameister, H. *Trends Genet.* **20**, 598–603 (2004).
9. Jensen, L. R. *et al. Am. J. Hum. Genet.* **76**, 227–236 (2005).

Die hard

Grenada's monkeys have thrived despite a huge degree of inbreeding. But now they face a new genetic bottleneck in the wake of Hurricane Ivan. Sharon Levy investigates.

When Mary Glenn returned to Grenada this January, she was stunned to see her study site, once a rich forest with a tall canopy of centuries-old trees, reduced to splinters. Glenn, a primate researcher and anthropologist, has spent more than a decade in the West Indies studying the mona monkey (*Cercopithecus mona*) in Grenada's highlands. But she was at her home in California when the full force of Hurricane Ivan hit the island on 7 September last year, flattening the forest and leaving a maze of stumps and twisted logs. Even though she'd seen the disaster on the news, the reality of the destruction was shocking.

The forest wasn't the only casualty. The hurricane tore through the island's towns, ripping off roofs and sucking glass clean out of windows. Many houses were reduced to their foundations, leaving the locals to camp out among the wreckage of their former homes.

Although the islanders' situation has touched Glenn deeply, the plight of the monkeys is also close to her heart. The storm killed much of the forest that the monas depend on for food, and even those trees left standing were stripped bare of their leaves and fruit. Hungry monkeys have been driven into the open to loiter at the edges of towns, raiding kitchens in a desperate search for food. It's likely that the majority of the island's 6,000 or so monkeys died during the storm, says Glenn, and more could perish in its aftermath. As the islanders set about rebuilding their homes, can the monkey population also find the strength to recover?

It's possible: the monas have a long history of survival. Grenada's monkeys are descended from a tiny population — perhaps



Before the storm: Mary Glenn (right) has spent the past decade studying the genetically isolated mona monkeys (above) in the West Indies.

a single pregnant female — brought from the West African island of São Tomé sometime in the eighteenth century by sailors on the slave trade route. As a result, today's mona population is highly inbred and displays practically no genetic diversity. When Glenn recruited a geneticist to analyse the monkeys' maternally inherited mitochondrial DNA, she kept checking and rechecking her work, fearing that she had accidentally rerun a sample from the same individual. "They are as alike as identical twins," says Glenn.

Yet before the hurricane struck, the monas were in rude health — not necessarily what one might expect from a highly inbred population. Given that, Glenn is optimistic about their chances in the face of yet another genetic bottleneck. "The monas will bounce back," she says.

Single life

Periods of genetic isolation have played an important role in the evolutionary history of most cercopithecine monkeys, commonly known as guenons. At least 21 species and more than 60 subspecies of guenon have been identified in Africa, ranging from Sierra Leone in the west to the headwaters of the Nile in the east. The monkeys come in a wild array of brilliant face and body patterns, which evolved when ancestral populations were isolated in patches of forest during droughts long ago. Guenons attract the interest of evolutionary biologists in part thanks to the monkeys' ability to mate and produce viable offspring with other species of the genus. Cross-species guenon pairs with wildly different



physical traits and even different numbers of chromosomes have been known to produce healthy young.

Glenn, now a professor at Humboldt State University in Arcata, California, first came to Grenada as a graduate student in 1992 to study the guenons. The thriving population of genetically isolated monkeys seemed to be the perfect natural laboratory, particularly as many of the monas were relatively fearless and easy to observe. Before Ivan struck, Grenada was a mona paradise; the monkeys had few competitors for food, and their only predators were a small number of hunters who culled perhaps a few tens of animals each year.

Glenn spent the next three years roaming the forest of the Grand Etang Reserve, developing close relationships with the island's

Wrecked: Hurricane Ivan left a trail of destruction in its wake when it struck Grenada last September.

monkeys and people. At that time, almost nothing was known about the natural history of monas, so Glenn and her co-workers were able to document some previously unknown behaviour^{1,2}. She noted, for example, that the monkeys have a unique copulation call—a distinctive, multisyllabic cry.

Her work also revealed interesting social structures. Primatologists who had studied guenons in the past had found that they tended to live in clans made up of a dominant male, several females and their young. Adult males that lacked the clout to control a harem were thought to live solitary lives. But the most obvious groups of monas in the Grand Etang Reserve are all-male gangs that hang around the reserve's visitor centre, scamming tourists for donations of fruit. "Monas are the only guenons that have documented, stable, very social, all-male groups," says Glenn. "They groom each other, sit in contact, engage in homosexual sex, even share food."

Monkey puzzle

But perhaps the biggest surprise for Glenn about Grenada's mona population was the fact that the island monkeys seemed to be healthier than their mainland counterparts. "I expected to find problems and quirks in the population," she says; this often happens in inbred populations, as rare recessive genes pair up and are expressed. But she found none. "I think it was just chance that a very healthy female came to Grenada, that she didn't have a lot of deleterious genes,

and that the island had no environmental factors that happened to set off the genetic weaknesses of her descendants," says Glenn.

The damaging effects of inbreeding are well documented in captive animals. There is a correlation between a decline in genetic diversity and an increase in malformed or stillborn offspring, and some studies suggest that the efficiency of the immune system also declines³. Statistically, inbreeding makes a population more vulnerable to a freak storm or new illness. But some species have done quite well with a minimal genetic repertoire. The Mauritius kestrel (*Falco punctatus*), for example, was decimated by the impact of pesticides and dwindled to a single breeding pair in 1974. Yet today there are at least 200 healthy pairs living in the wild. And, as scientists as early as Charles Darwin noted, tiny founding populations on islands can do amazingly well in the absence of disease and other challenges.

Despite such examples, conservation biologists still often say that once a population declines below a certain level there is no point trying to rescue it, as it will be doomed by the effects of inbreeding. "I heard repeatedly from other biologists when I was working on the California condor project that it was a total and monumental waste of time and money," says Keith Bensen, Glenn's husband and a wildlife biologist at Redwood National and State Parks in California. "Well, it's not." California condors (*Gymnogyps californianus*) were decimated by hunting and habitat loss, but a captive breeding programme begun

with a handful of survivors helped reintroduce them to the wild. "Sometimes the population will survive and thrive," Bensen says.

But Glenn cautions that these success stories don't mean that all small populations will fare equally well. Many of mainland Africa's guenons have been driven into small 'islands' of habitat by heavy logging, but unlike Grenada's monas, these monkeys are also targeted by heavy hunting and are exposed to competition and disease from other primates. "We're not advocating the idea that it's OK to isolate monkey populations in west Africa," says Glenn. "You can't take them down to one or two animals and assume they can recover." Although a lack of genetic diversity doesn't always spell doom, these extra pressures can tip some species into extinction.

Climbing back

Back on Grenada, the monas that lived through the initial chaos of Hurricane Ivan are continuing their struggle for survival. The quest for food has forced them near to houses and roads, and several have been run over by cars.

But people have reported seeing monkey mothers with their young—a promising sign of recovery. And a few palms are beginning to grow leaves and form new fruit. Officials on the island have also imposed a temporary ban on monkey hunting. Glenn, whose most recent trip to the island was focused on bringing material and financial aid to families and community organizations, hopes to return to Grenada soon to study the monas' responses to life after the disaster.

The hurricane's long-term effects will make this a difficult task. Four months after the storm, the Grand Etang Reserve looks green once again. But instead of a 45-metre-tall canopy of old-growth forest, there is a landscape of stumps and shattered tree trunks, draped in vines. The rains have come to Grenada, and with most of the trees dead, the rugged terrain is collapsing. A growing number of landslides now darken the streams that are the island's source of drinking water, and dangerous terrain makes it nearly impossible to move through the forest. Park workers plan to begin replanting trees in the reserve to stop this decline.

It will take years for the land and the monkeys to recover. And it will take several generations of monas—some 20 years or more—before Glenn can truly assess the effect this latest genetic bottleneck has had on the population. Like everyone else on the island, Glenn and the monas now need to start over. ■

Sharon Levy is a freelance writer based in Humboldt County, California.

1. Glenn, M. E. *Am. J. Primatol.* **43**, 167–173 (1999).
2. Glenn, M. E., Matsuda, R. & Bensen, K. in *The Guenons: Diversity and Adaptation in African Monkeys* (eds Glenn, M. E. & Cords, M.) 133–146 (Kluwer Academic/Plenum, New York, 2002).
3. Acevedo-Whitehouse, K., Gulland, F., Greig, D. & Amos, W. *Nature* **422**, 35 (2003).

NIH must tell whole truth about conflicts of interest

Little has yet been disclosed about a consulting spree that has left a legacy of harm.

Sir—Your Editorial “Taking a hard line on conflicts” (*Nature* 433, 557; 2005) understates the extent of the embarrassing failure by the US National Institutes of Health (NIH), during the past year, to disclose information about NIH scientists serving as paid consultants to private companies. Having been a scientist at the NIH since 1967, as a physician, a cell biologist and now a science administrator, I must add that the habit of non-disclosure continues, making further embarrassments likely.

In November 1995, the then-director, Harold Varmus, made an unpublicized change to the rules on paid consulting by NIH scientists. The permissiveness of the 1995 rules opened the way for a spree of consulting that ended on 1 February 2005, when a stricter policy was announced. The previous decade left a legacy of harm, only part of which has come to light.

In articles published since December 2003, the *Los Angeles Times* has given ten specific examples of NIH scientists with financial conflicts of interest. However, details of hundreds of others remain hidden, and the extent of the damage caused since 1995 is unknown.

NIH director Elias Zerhouni told a Senate subcommittee last year that he had found no harm to patients as a result of NIH scientists’ outside financial arrangements. However, he provided no evidence to support this claim and did not explain how he arrived at his conclusion.

On 2 February 2005, I attended a large meeting of NIH scientists at which Zerhouni spoke of “very disturbing” product endorsements by NIH scientists “speaking for a company on behalf of a product to entice physicians to prescribe a product at greater levels”. He provided no

information about the products, the companies, the payments or the scientists.

During the past year, Zerhouni has said that achieving transparency (defined by the NIH as public availability of financial information filed by NIH employees) is “absolutely critical” and “first on the agenda”. But much of the information about past consulting, going back to November 1995, remains undisclosed.

The NIH’s defensive approach — one congressman called it stonewalling — has proved a disastrous miscalculation. If the NIH does not unflinchingly seek the facts and release them, they may come out anyhow in other ways. Then the NIH’s reputation, already at the lowest point in its history, will suffer further.

Ned Feder

National Institutes of Health, 2 Democracy Plaza, Bethesda, Maryland 20817, USA

Limits to growth may be subtle but still inexorable

Sir—Dick Taverne, in his review of James Gustave Speth’s recent book *Red Sky at Morning*, comments about a “wildly inaccurate” 1972 report by researchers at the Massachusetts Institute of Technology (MIT): “*The Limits to Growth* predicted that we would run out of gold, zinc, mercury and oil before 1992” (“When greens see red”, *Nature* 432, 443–444; 2004). This is a very common myth which needs correction.

The precipitous collapse of economic growth projected by the MIT report does not occur before 2020 in any of the model runs, so it is hard to see how its doomsday projections have been disproved. For a 30-year-old projection using some eight state variables, the MIT model did a surprisingly good job of predicting where we were in 2004, as an update shows (D. A. Meadows *et al.* *Limits to Growth: 30 Years On*, Earthscan, London, 2004).

One run of the model even had the natural-resource constraint removed by assuming an unlimited source of energy.

The model’s mechanism of collapse is more subtle than simple resource exhaustion. Suppose, as an example, that when the Chinese want to replace the steel mills they are now building, they find that materials are available but costs are 10 times higher, while their productivity in the interim has only doubled. Hard-pressed to keep steel production growing at all, they will have even less mill capacity available to make materials for replacement in following

years, and so eventually will be overwhelmed by the spiralling legacy of repairs.

Whether or not this is a model for a reckless world’s future, failure of revenue to cover accelerating physical depreciation is a frequent explanation for urban decline.

David Fisk

Department of Civil and Environmental Engineering, Imperial College London, London SW7 2AZ, UK

More power needed to probe cloud systems

Sir—By using the distributed power of personal computers around the planet, D. A. Stainforth and colleagues (*Nature* 433, 403–406; 2005) have quantified uncertainty in forecasts of global warming resulting from a doubling of CO₂. If warming occurred at the upper end of the predicted range, the effects for humankind would be utterly catastrophic.

Such forecast uncertainty arises, in large part, from the way cloud systems are represented in existing global-climate models. Because of limitations in computer power, the underlying laws of physics cannot be used directly to simulate such systems; instead they are represented by approximate formulae with uncertain parameters. In the paper by Stainforth and colleagues, an ensemble of many thousand individual global-warming forecasts was analysed. The forecasts were all made with the same climate model. However, the parameter values varied from one forecast

to another — the intra-ensemble variation in these values being consistent with their inherent uncertainty.

Global-climate models are now being developed in which one of the most important types of cloud system, ‘organized deep convection’, is computed explicitly. Such convective systems, with horizontal scales of tens of kilometres, play a key role in climate: they cool the Earth’s surface, they transport water from the surface into parts of the troposphere where they can contribute to the greenhouse effect, and their kinetic energy influences global-scale climate circulations.

Unfortunately, even on the most powerful computers, climate simulations in such models take nearly as long as real time. Such high-resolution climate models cannot be run efficiently using distributed computing technology.

To reduce uncertainty in estimates of global warming, the climate-modelling community requires substantially more powerful computational resources than are currently available, so that more of the climate system can be simulated directly from the known laws of physics.

In view of the seriousness of the global-warming problem, plans for developing such a facility for climate prediction should occur at the international level, so that national resources can be pooled and scientific collaborations enhanced. Within Europe, the European Union could play a leading role.

T. N. Palmer

European Centre for Medium-Range Weather Forecasts, Reading RG2 9AX, UK

Just for fun?

Working out why animals play is no easy task.

The Genesis of Animal Play: Testing the Limits

by Gordon M. Burkhardt

MIT Press: 2005. 518 pp. \$50, £32.95

Bernd Heinrich

A kitten batting a ball of yarn, kids on a swing, or an adult wielding a fishing-rod — few would disagree that these behaviours can be described as play. Yet in the study of animal behaviour, the phenomenon of play is an anomaly. It is said to be adaptive and yet it involves the expenditure of much energy, often with no apparent pay-off. When a certain behaviour is found to have obvious pay-offs or functions it is, almost by definition, no longer 'play' but is defined by its function, such as foraging, predator avoidance or mating.

According to the simplest, most short-term definition, play is 'just for fun'. But in the long term it may also be practice for a future role, although the ultimate pay-off may only be determined over a lifetime. Consider the activity of batting a ball around. That's play, isn't it? But if someone got paid for it, would it still be play? Even without pay, it may be for practice. And what about others who expend time and resources to watch this senseless activity? It would be difficult to assign adaptive value to these behaviours, or to measure them objectively. Should play then be defined by internal motivation — pleasure, fun? But the motivation of internal pleasure can apply to many complex behaviours, such as hunting, birdsong, sex or chasing a frisbee, although only the latter is likely to be called 'play'. Is play then only 'senseless' behaviour, or is it simply behaviour for which an ultimate function has yet to be discovered?

Play may be plagued by paradoxes and enigmas, but it is a genuine behavioural phenomenon. It is an appropriate subject for enquiry, if only because we know so little about it, despite the interest of scholars who for centuries have tried to define it, fix its boundaries and fathom its functions. For the most part there has been little progress — instead, the subject has become entangled in a web of definitions and controversies. It is clearly time to re-examine play.

The Genesis of Animal Play does not really explore the limits that I allude to above, but to date it is the most comprehensive and illuminating effort to come to terms with this enigmatic topic. Even though Gordon Burkhardt claims his book "is not meant to be a thorough review of play research in animals or people on either a narrow or broad scale", I believe nevertheless that it does

Dolphins are thought to engage in play-fighting — but is it just for fun or does practice make perfect?

come close. Burkhardt reviews the literature (about 1,300 references are cited) and refers to most of the serious attempts to study play. In his attempt to understand its origins and nature, he incorporates both comparative and multidisciplinary approaches.

In the first part of the book, Burkhardt explores the diversity of play behaviour and the history of various theories, definitions and controversies surrounding it. He also proposes criteria for a modern definition of play. He champions the view of ethology laid down by Niko Tinbergen, which says that any definition of a behaviour must encompass four entirely different types of problem: its causation or mechanisms; its adaptation or function; its development or ontogeny; and its evolution and phylogeny. Burkhardt adds a fifth category: the world of private experience. Researchers using this last criterion would ask whether all play is accompanied by one or a few specific emotions (presumably after play has already been identified, not to identify it).

In the second half of the book, Burkhardt examines the phylogeny of play, reviewing examples from studies of mammals, birds, reptiles, fish and invertebrates. Most of the examples of play in the literature come from a small number of placental mammals, in which play is clear-cut. However, because the focus of the book is the origin and function

of play, the most relevant examples are those at the boundaries where play is not readily distinguished from non-play. Even among the mammals and the few birds studied, there is a great variety of play or potential play behaviour, so ecological, social and other potentially relevant factors may shed light even within this narrow taxonomic grouping. At the borders, in fish and invertebrates, descriptions of putative play behaviour remain anecdotal. I suspect that few of these will sustain the five criteria for play that the author sets up, but he remains open-minded to the possibility.

Burkhardt concludes by strongly backing the 'surplus resources' theory as a way to predict where and when we might expect to find playful behaviour. This idea is an elaboration of one proposed in 1795 by the German playwright Friedrich Schiller, who wrote: "An animal may be said to be at work when the stimulus to activity is some lack, and it may be said to be at play when the stimulus is sheer plenitude of vitality." Schiller's idea was further elaborated by Herbert Spencer in 1872. Burkhardt brings it into the social context and adds the adaptive significance — that play not only originates from, but also creates, surplus resources that may be useful on subsequent occasions. I am not sure how or at what point in the life of an animal such surplus resources would

manifest themselves or how an ethologist could demonstrate their existence. However, the idea surely applies to the activity of scientific research, and perhaps even to the writing of a book. ■

Bernd Heinrich is in the Department of Biology, University of Vermont, Burlington, Vermont 05405, USA

To infinity and beyond!

Parallel Worlds: The Science of Alternative Universes and Our Future in the Cosmos

by Michio Kaku

Allen Lane/Doubleday: 2005. 448 pp.

£20/\$27.95

David Lindley

Michio Kaku is an enthusiast. In the breathless introductory chapters of *Parallel Worlds*, he speeds through the history of cosmology and fundamental physics to persuade us that science has never been as thrilling as it is now. Edmund Halley may have been “shocked beyond belief” and “staggered” by Newton’s calculation of planetary orbits, but that’s small potatoes compared with the staggering science that goes on nowadays. String theory is staggering; quantum computers are truly staggering; the recent discovery of accelerating cosmic expansion left scientists dumbfounded; their minds reel at the idea of parallel universes. And when they’re not reeling and staggering, they have to contend with the “insane conclusions of quantum theory”.

Melodrama aside, Kaku’s book provides a sprightly and generally efficient account of our present understanding of the Universe, in which astronomical observation and elementary physics combine to give a coherent, if rather bizarre, picture of the cosmos. Plain ordinary matter, such as we are made of, is tame stuff. The Universe is mostly dark matter and dark energy, both revealed by the dynamics of galaxies. As yet there is no convincing explanation for either.

But string theory will take care of that. Beginning with the now familiar idea that all particles are in fact little wiggly loops inhabiting extra dimensions, Kaku sketches how gravity may one day be unified with quantum mechanics. Then we will understand space-time and all the matter and forces in it. There was a time when string theory seemed unique: only one version of it could possibly work. But more versions appeared, to be subsumed some years later into M-theory, which unifies string theories at the cost of introducing fundamental surfaces of higher dimensions (called ‘branes’).

It is a little unsettling now to learn that

Brane teaser: if parallel universes exist, they might take the form of bubbles in different dimensions.

physicists are drowning in branes, as Kaku puts it, and that they are beginning to think that M-theory itself must be part of some yet more basic theory that is awaiting discovery. This is progress, Kaku assures us.

Although he has every confidence that the quest for a theory of everything will succeed, perhaps in just a few decades, Kaku is vague about what success will look like. He describes how observations of gravitational waves or of temperature fluctuations in the cosmic microwave background, or perhaps experiments at CERN’s upcoming Large Hadron Collider, might reveal tiny signs of M-theory seeping from higher dimensions into our own poor world. Then again, they might not, in which case we have to fall back on neopythagorean arguments of pure mathematical consistency to let us know which is the right theory. If that theory is unique, we would “know the mind of God”, says Kaku, borrowing a favourite phrase of Einstein’s. Or there could be lots of legitimate theories, so that our Universe would be merely one of many possible universes. (I suppose this means that even God couldn’t make his mind up.)

Whatever universe we live in, it is bound to end badly. Stars die and cosmic expansion can only lead to a cold, utterly attenuated emptiness. But we can escape this dreary fate, Kaku says. Once physicists have figured out the theory of everything, they will be able to manipulate space-time to create a wormhole through which we can pass to a new universe and start over again. Or maybe, if it is too difficult to send our physical selves to another universe, we can capture our

civilization’s entire sum of information and convey it down the wormhole in the memory of a quantum nanobot. This is the big payoff: by fully understanding physics, intelligence can escape its cosmic doom.

Parallel Worlds slides from even-handed discussion of what physicists and cosmologists are actually doing into areas that are, shall we say, a touch speculative. Kaku mentions problems and difficulties along the way, but he gives the distinct impression that the prospects described in his final chapters are, if not exactly within our grasp, at least reasonable extrapolations. Some readers will no doubt be entranced by the grand vision, the great sweep of imagination. For me, the parade of mind-boggling ideas breezily portrayed here — the Big Bang as a collision of branes, time travel around massive rotating cylinders, matter as holographic projection from higher to lower dimensions — created a different cumulative impression. The great enterprise of fundamental physics began to seem larky, whimsical and not quite serious.

String theory and its cosmological implications embody, for the moment, a research programme much more than a genuine theory of physics. Getting from one to the other demands detailed and diligent effort. There is no guarantee that it will all work out as desired. But Kaku takes the truth of the big ideas more or less for granted, and races ahead to the most far-reaching and literally cosmic conclusions. This is intellectual entertainment — brane candy, if you like — but will it ever be more than that? ■

David Lindley is a freelance science writer in Alexandria, Virginia, USA.

A chemical conspiracy?

The Fluoride Deception

by Christopher Bryson
Seven Stories Press: 2004. 272 pp.
\$24.95, £17.99

James Clark

Fluorine chemists are proud of their unique and often unpredictable element. It sits at the top right-hand corner of the periodic table above its domain of conquests — the range of compounds it forms is remarkable, with few elements escaping its extreme reactivity and the strong bonds that it forms. The range of fluorine compounds is extraordinary, from the extremely inert calcium fluoride (the mineral fluor spar) to the highly aggressive hydrogen fluoride. This versatility has been used to great effect in applications as diverse as non-stick cookware (Teflon), high-temperature insulation (sulphur hexafluoride) and many of the most active pharmaceuticals and agricultural chemicals on the market.

Fluorine has also been a key player in several more controversial areas, although fluorine itself is seldom the culprit. It was vital to the Manhattan Project (through the use of uranium hexafluoride for isotope enrichment) and it is the major element in CFCs (chlorofluorocarbons), but it is not responsible for their destructive effects. More recently, the remarkable stability of heavily fluorinated molecules has been shown to be an environmental risk as a result of their persistence in the environment.

The uses of fluorine that have consistently received most attention in the media and from the public are the fluoridation of water and fluoride dental treatments. The arguments have raged for more than forty years, and in *The Fluoride Deception*, Christopher Bryson raises the stakes by reporting a great

Clear risk: fluoride levels only slightly above those used in tap water could be harmful.

deal of relevant and often alarming research, and by telling a series of human stories.

Bryson gives detailed consideration to the scientific literature on the biological actions of fluorides and the major debates about their use. Some of the findings he has identified are very interesting, such as an article from researchers at the University of York showing that water fluoridation may be responsible for reducing the incidence of cavities by only 15% — a far cry from the much higher figures suggested when the fluoridation campaign was at its peak in the 1960s. A claim in the April 1944 edition of *Time* magazine was that it led to “perfect teeth”. The evidence for the harmful effects of fluoride at only slightly greater concentration than artificial

fluoridation levels is disturbing, especially when so much of the research data have been widely available for many years. Some of this apparently damning evidence was aired at a lecture to the US National Institutes of Health more than 10 years ago.

Bryson also makes several claims about the suppression of antifluoridation evidence, the discouragement of related research (for example, through the withdrawal of federal funds) and, most controversially, that fluoridation was a direct consequence of a group of powerful industries seeking to solve the costly problems they had in dealing with fluoride pollution. The nuclear, aluminium and fertilizer industries (which are large-scale users or producers of fluorine compounds) are subject to Bryson's close scrutiny and criticism. In 1975, the US government estimated that 350,000 people in 92 different occupations were being exposed to fluoride in the workplace. He suggests that the decision to fluoridate the public water supply was made with incomplete test data and was driven by the government as a means of disposing of large quantities of fluoride waste. Bryson's attempt to link industry pressure to the decision to begin fluoridation, from the Manhattan Project onwards, is the most intriguing aspect of his book.

Bryson shows skill as a writer and adds colour to his account with many alarming and well-referenced stories, often focusing on individuals. Unfortunately, his desire to make the book more exciting leads him into the all-too-familiar trap of tarring with the same brush anything associated with, or even sounding like, fluoride or fluoridation. This is especially ironic after he starts the book with “notes on terminology”, saying “fluorine and fluoride should not be confused”. However, in the next section he tells us “the same potent chemical that is used to enrich uranium for nuclear weapons, to prepare sarin nerve gas ... is what we give to our children”. No doubt we can expect a series of books on chlorine (“the same potent chemical used in insecticides is what we put on our children's food”) and oxygen (“the same potent chemical used in the strongest acids is what we allow our children to breathe”).

No chemist would dispute the extreme hazards of many fluorine chemicals, but to group all fluorine chemicals together as “bad” is wrong. The book is peppered with similar absurdities, which will be annoying to those who know their chemistry but dangerously misleading to those who don't. There are other errors of identification and labelling, such as the description of Unilever as one of the “world's most powerful drug companies”, leading to further confusion. These flaws detract from what is otherwise a thought-provoking and worthwhile book. ■

James Clark is at the Clean Technology Centre, Department of Chemistry, University of York, York YO10 5DD, UK.

New in paperback

Platypus: The Extraordinary Story of How a Curious Creature Baffled the World by Ann Moyal (Johns Hopkins University Press, \$16.95, £11.50). “Her book has something for everyone — the excitement of a detective story, the history of biological ideas, the frustration at the fact that morphologists appear not to want to understand the evolution of structure in relation to function.” David Penny, *Nature* **412**, 857 (2001).

The Great Influenza: The Epic Story of the Deadliest Plague in History by John M. Barry (Penguin, \$16). “Barry writes like an angel. He totally immerses us in the 1918 flu battle.” John Oxford, *Nature* **429**, 345–346 (2004).

Critical Mass: How One Thing Leads to Another by Philip Ball (Arrow, £9.99). “For anyone who would like to learn about the intellectual ferment at the surprising junction of physics and social science, *Critical Mass* is the place to start.” Steven Strogatz, *Nature* **428**, 367–368 (2004).

Hitler's Scientists: Science, War, and the Devil's Pact by John Cornwell (Penguin, \$16, £8.99). “Cornwell has written an engaging synthesis of the original research ... a useful compilation for readers who would like just one volume on science under the Nazis.” Kristie Macrakis *Nature* **425**, 766–767 (2003).

Schrödinger's mousetrap

Part 9: Out in the cold.

Peter Tuthill

The rain spattered on the glass panes of the conference-hall foyer as Lister shuffled through the last pages of the partially comprehensible report from forensics. He cast a withering glance at the expensive low-fat macchiato, untouched on the table by his elbow. On another day, he might have taken up his personal crusade for normal coffee, available to those with no fluency in Italian. The festive curly star that the girl had cleverly made in the foam on top twinkled back, taunting him.

In a perverse corner of his mind, he was delighted as the rain strengthened, engorging the branching braids of water on the window. Somehow it never felt right conducting a murder investigation on a sunny day, and he revelled in today's sympathetic backdrop. For murder it clearly was, any cobwebs of doubt now dispelled by the report in front of him.

He let the new information sink in: Rufus Jaeger died with his stomach laced with a cocktail of anaesthetics and paralytics. Was the murderer making absurdly sure, or was this another tangent, leading into the twisted corridors of academic intrigue?

With a sigh, Lister admitted defeat in the glaring match with his macchiato and, taking a tentative sip, pulled a face before suddenly lowering the cup. Outside the double revolving glass doors was Wilfred de Bruijn, ten minutes early for his interview. Recognizable more by gait than appearance, he moved like a small determined bear, bent forward and seemingly oblivious to the world around him. With increasing puzzlement, Lister watched de Bruijn pass the doors, hunch his coat against the rain, and skirt round the building, entering instead through an unnoticed side-door after an awkward struggle with its latch. As he came past the cafe tables, shaking raindrops from his wild hair and muttering to himself, Lister remarked dryly: "Quite a performance, Dr de Bruijn, but tell me, who are you so keen to keep ignorant of your presence here today?"

Startled suddenly from his internal world, de Bruijn stammered: "Prof ... er ... Inspector Lister, I was just coming to the interview. I hope I am not late."

At Lister's continued quizzical look, colour rapidly rose from his neck to his ears, and he continued. "You mean the ... er ... door? Well, I have nothing to hide and am not avoiding anyone." Here the squirming discomfort became almost palpable. "I ... er ... never go



through revolving doors. Ask anyone from the lab. I just don't like them."

As they made their way to the room commandeered by the police, Lister mused on the contradictions that made up the man beside him, with his intense coal-black eyes framed by thick eyebrows like animated hairy caterpillars. On one hand, this was by all accounts a man of fierce intellect, able to revel in arcane quantum physics. Yet on the other, a man too superstitious to walk through a revolving door. Lister wondered if his mind might be likened to those quantum states he had been reading about after his interview with Dubois. Depending on context, de Bruijn's mind was in an irrational or rational state. Lister drained his coffee, clearing his head of metaphysics — somebody should have warned him about spending too much time among physicists.

The interview was more a stream-of-consciousness diatribe and it was all Lister could do to channel the rant. The traces of South African were almost polished away under a studied BBC accent, but the bright kid who'd clawed his way out from small-town Transvaal was still there, trapped in the cage of his own ambition and raging against a world where intellectual ability was neither necessary nor sufficient to make it to the top.

"Rufus was a scientific parasite, Inspector Lister, a showman, politician and front-man, but he never had an original physical insight in the years I spent working for him. And the minions who put in all the ideas and work just got walked on. Last month, accepting €100,000 prize money for an experiment designed by me that he barely even understood, the use of the scientific royal 'We' in

his speech was the only acknowledgment anyone else got. Well, except..."

"Except?" prompted Lister.

"That damn Ludmilla princess. She's gorgeous and she knows it. I have five times her publications, but I get recognized with 'Ahh — you're that guy who works with Ludmilla!' We may know maths, Inspector Lister, but we are still apes."

"And you and Ludmilla..."

De Bruijn gave a bitter laugh. "No, Inspector, you are in a blind alley there. I think Ludmilla knows to bestow her favours where she stands to gain the most." He glanced pointedly at Jaeger's file on the table. "Although I tried warning her to keep a professional distance."

"Did she ever reply?" Lister asked casually. De Bruijn shook his head.

"Witnesses report an altercation — can you clarify what you were doing during the coffee break?"

"I first ran into that incompetent upstart Jirong Feng, who was in my face about something. Then Nigel Lorimer said he urgently needed to talk to me. But if you must know, I had been meaning to speak to Rufus the whole day. That was my damn mousetrap — well, except for the stupid cheese and fairy-lights. And guess who was sitting back in row 7? Everyone else was up on stage, even that brain-dead yes-man Trotman who nearly burned the lab down with the laser last year. And what the hell are we doing with an industrial laser able to punch holes in steel? So after I got rid of Nigel I confronted Rufus, who was talking to Ludmilla. He dismissed me with his typical arrogance: 'Run along, Wilfred, if I want your opinion I will ask. Don't go getting ideas above yourself.' And he turned his back. After that, I needed a long walk outside to calm down."

Lister drifted down to the foyer, his mind whirling with possibilities. Overhearing animated banter on whether the appropriate collective noun for physics professors should be 'a guffaw', he resisted the urge to chip in with 'an asylum'.

Suddenly, a light came into his eyes as the pieces fell into place. Lister reached for his mobile, he had two important calls to make. "Lister here. Get me forensics."

To be continued...

Peter Tuthill is in the Astronomy Department, School of Physics, University of Sydney, Sydney, NSW 2006, Australia.

Who do you think killed Rufus Jaeger? Catch up on all the evidence and vote for your suspect at
www.nature.com/news/mousetrap

She moves in mysterious ways

Chris Gunter

The human X chromosome is a study in contradictions. The detailed sequence of the X, and a survey of inactivated genes in females, help to illuminate this unique 'evolutionary space'.

The X chromosome was originally named X for 'unknown', remaining an oddity that has puzzled geneticists for centuries. It is the only chromosome to have one of the pair inactivated in one sex (females), and it is riddled with repeat elements, making it especially tough to produce a detailed gene sequence. From the papers by Ross *et al.*¹ and Carrel and Willard² (pages 325 and 400 of this issue), we learn that the wait was worth it.

The mammalian X and Y chromosomes arose from one pair of autosomes (that is, all the other chromosomes) 300 million years ago. Female mammals have two Xs; males have an X and a Y. After the X and Y seceded, mutations in genes on the Y made it the male-determining chromosome, and the pair began to diverge. The X and Y still share a few genes, mainly in the 'pseudoautosomal' region on the tip of the X; here the two chromosomes still exchange DNA to maintain proper segregation in cell division. Over time, the Y disintegrated to a shadow of its former self — but as long as the genes for maleness are preserved, most of the formerly autosomal genes seem to be largely extraneous because men already have one copy on the X.

In turn, the X developed a way to inactivate — silence — most of the genes on one of the two Xs in females, so that males and females would in large part have the same dosage of gene products. Early in female development, cells randomly choose either the maternal or paternal X to be the active X chromosome. The other one then transcribes large amounts of a large RNA from a gene in the middle of the chromosome, *XIST*. Only this chromosome becomes coated by *XIST* RNA, and thereby silenced by modification of its DNA and associated proteins. This choice is permanent, and has certain consequences. A famous example is the calico cat (Fig. 1). Similarly, human females are mosaics of the X chromosomes from each parent, and the severity of an X-linked disease in a female depends on the percentage of the cells in which the mutant gene concerned is silenced or expressed.

The inactive human X remains silenced even after being transferred into mouse cells, allowing for the formation of hybrid cells containing the inactive but not the active human chromosome. Carrel and Willard² have exploited this system to create a near-



Figure 1 X-chromosome inactivation made manifest. The coat colours of calico cats, such as that pictured here, vary in individual animals depending on which X was inactivated in different cells during early development.

complete catalogue of the inactivation state of X-linked genes. They find that 75% of genes are permanently silent, and about 15% permanently escape inactivation, meaning that they are expressed at twice the level in females as males. The remaining 10% are expressed in some inactive Xs but not others, indicating variability in human females that is likely to have medical relevance. Having twice the amount of any gene product could result in female–female or female–male differences: the possibilities, as yet undefined, are intriguing.

The mechanism of X-chromosome inactivation remains under debate, but the upshot is clear: in evolutionary terms, according to a principle known as Ohno's law, inactivation essentially froze the genes in place except in the pseudoautosomal regions. Yet comparative analyses of vertebrate genomic sequences tell us that human chromosomes have engaged in a big evolutionary swapping exercise, exchanging whole segments of DNA between them, even on the X and Y. The complete sequence of the gene-containing parts of the X and Y chromosomes^{1,3} has enabled Ross *et al.*¹ to construct a remarkably detailed portrait of the X's past.

Previous comparisons between X- and Y-shared genes suggested that there are five evolutionary strata on the X, starting at the

bottom of the long arm of the chromosome⁴. These strata represent large chromosome chunks formerly shared by X and Y — some original to the autosome pair, some arriving later — until inversions on one of the pair meant that X and Y could no longer exchange DNA in that region, moving the pseudoautosomal boundary upwards with each inversion. Selective pressure presumably maintains the small pseudoautosomal region that remains, so any further erosion of shared genes should be minimal.

X-chromosome inactivation is then proposed to work on each newly non-exchanging stratum, silencing genes to maintain proper dosage, while the Y version degrades over time. Indeed, Carrel and Willard² found that the tendency to escape inactivation correlates with age of stratum, so that fewer genes escape in the oldest and more in the youngest. Despite all the shuffling, the X-inactivation centre remains at about the middle of the chromosome, also probably owing to selective pressure.

What genes have remained on the X? Because males have only one copy, diseases resulting from mutated X genes are often easier to identify in males — hence the early discoveries of genes involved in haemophilia and muscular dystrophy. According to previous dogma, the X is low in genes implicated

R. PLANCH/NEHA

in tumour formation and growth but high in those related to sex and reproduction⁵. It is surprising, then, that Ross *et al.*¹ predict that a full 10% of protein-coding genes on the X produce 'cancer-testis antigens', a class of genes normally expressed mostly in the testis but whose activity is increased in testicular cancers, melanomas and other cancers.

These proteins are key targets for the development of cancer vaccines, as they are usually seen by the immune system only when they occur in tumours, and their expression is linked to the same mechanisms at work in X inactivation. As we know little about their normal function⁶, tailoring treatments will require a much better understanding of the role of the X in cancer. Given the prediction⁷ that genes on the X can benefit males even at the expense of females, as the detrimental effect would be masked in females, it may be that the genes encoding cancer-testis antigens convey some selective advantage to males, allowing some of the gene family to expand independently in both humans and mice¹.

What about the sequences between the genes? The X chromosome is very rich in repeats (56% compared with a 45% average for the whole genome), and 14 gaps remain in the estimated 1.5-million base-pair sequence — despite the intensity of the sequencing effort¹. But it's not just any repeats that matter. A whopping 29% of the chromosome consists of DNA repeats known as 'long interspersed nuclear elements' (LINEs), thought to be 'selfish' bits of DNA that jump around the genome. The main LINE constituents concerned are members of the L1 family and, in theory⁸, L1 repeats may serve as 'booster elements' for the X-inactivation signal, which emanates from around the middle of the chromosome with the transcription of *XIST* and spreads down the chromosome arms.

In support of this theory, the sequence shows almost no L1 elements in the *XIST* locus itself, but high concentrations on either side. L1 sequences are more common in the older areas of the chromosome, with more genes subject to inactivation, and less frequent in the newer areas, where more genes escape^{1,2}. So it is possible that more genes may become subject to inactivation over time, as still-active L1 elements invade in higher numbers.

The exact mechanism for L1 involvement is not known: is this 'junk DNA' crucial in shaping the X, or merely the footprint left behind as the process progresses? Mice and cows do not show similar patterns of LINE elements on their Xs, or of any other known repeat element that might fit the bill, yet they have many similar features of X-chromosome inactivation. Factors involved in inactivating the X are modifications to the DNA and proteins in the chromosome, and formation of a complex with *XIST* RNA and

presumably associated proteins. Intensive research is aimed at finding out how these modifications might have evolved, and how they now establish and maintain the inactivated state.

Like the rest of the human genome, the primary sequence of this chromosome is an advanced starting point for exploring the mysteries of evolution and development. Geneticists eagerly await genomic sequences from non-placental mammals, such as the platypus and opossum, so as to reconstruct

earlier, proto-X chromosomes. As the rock band U2 would say, the X moves in mysterious ways, and we've just been given a preview. ■

Chris Gunter is a senior editor based at Nature's Washington office.

e-mail: c.gunter@naturedc.com

1. Ross, M. T. *et al.* *Nature* **434**, 325–337 (2005).
2. Carrel, L. & Willard, H. F. *Nature* **434**, 400–404 (2005).
3. Skalecky, H. *et al.* *Nature* **423**, 825–837 (2003).
4. Lahn, B. T. & Page, D. C. *Science* **286**, 964–967 (1999).
5. Kohn, M. *et al.* *Trends Genet.* **20**, 598–603 (2004).
6. Spatz, A. *et al.* *Nature Rev. Cancer* **4**, 617–629 (2004).
7. Rice, W. R. *Evolution* **38**, 735–742 (1984).
8. Lyon, M. F. *Cytogenet. Cell Genet.* **80**, 133–137 (1998).

Planetary science

Picturing a recently active Mars

Victor R. Baker

Discoveries made with the High Resolution Stereo Camera on the Mars Express orbiter show that, as recently as a few million years ago, the surface of Mars was being shaped by flowing water, lava and ice.

Spectacular ground-based images and chemical analyses of ancient sedimentary rock formations¹ leave no doubt that Mars had a watery ancient past. These discoveries by NASA's Mars Exploration Rover Mission apply to the Noachian epoch of Mars history — the first several hundred million years of the Solar System, when the impact rate of meteors and comets was much higher than in the past 3–4 billion years². Image interpretation has long been key to revolutionary changes in the scientific understanding of Mars. Three papers in this issue^{3–5} add to that tradition, but also add to the new evidence that Mars has been active in the geologically recent past.

The first vidicon camera images and atmospheric data from the Mariner fly-bys of the 1960s led to the view that the planet had been continuously cold and dry⁶. Thinking changed radically during the 1970s, when pictures of higher resolution and global coverage, taken from the Mariner 9 and Viking orbiters, revealed river-valley networks and huge channels carved by cataclysmic floods⁷. Nevertheless, in the theoretical syntheses that followed during the 1980s and 1990s, any active hydrological cycling was generally limited to the Noachian epoch. This MIDDEN hypothesis ('Mars is continuously dead and dry, except in the Noachian') prevailed for two post-Viking decades, during which there were no further pictures at increased resolution. Advocates of the MIDDEN hypothesis have variously argued that Mars is poorly endowed with water, that it is inefficient at releasing water to its surface, or that most of its water was lost to space.

In the 1980s, NASA planned a Mars geoscience and climatology orbiter, but many scientists wanted its payload to be limited to non-imaging experiments. Some even

proclaimed that imaging science was an oxymoron, and that the only use of pictures was for public relations. Thus, a high-resolution stereo camera system was rejected for what eventually became the Mars Observer spacecraft. The miniature, non-stereo, high-resolution Mars Observer Camera (MOC) was eventually added to its payload, but Mars Observer was lost in 1993, and MOC did not arrive at Mars until the Mars Global Surveyor mission of the late 1990s.

Despite being limited to non-stereo coverage of very small areas, the phenomenally high resolution of the MOC images contributed to notable discoveries. These discoveries included very recent, globally dispersed hillslopes, with gullies created by water⁸, as well as associated lava flows and large channels carved by cataclysmic flooding⁹. All of these were less than about ten million years old¹⁰, as indicated by the densities of impact craters measured from the high-resolution images. These features and many others¹¹ were clearly anomalous in regard to the prevailing MIDDEN hypothesis.

The High Resolution Stereo Camera (HRSC) carried by the European Space Agency's Mars Express Mission is the direct descendant of the camera that was rejected by NASA for the original Mars Observer. Although slightly lower than MOC in resolution, HRSC produces colour and stereo images in the broad areal context necessary for the fully effective application of geological analyses (Fig. 1). Initial results from the camera¹² identified two major anomalies in regard to the MIDDEN hypothesis. First, crater counts for calderas on five major volcanoes show that there has been episodically active volcanism during the past 20% of martian history, up to only two million years ago; second, glacial activity occurred in the equatorial regions of Mars, at the base of

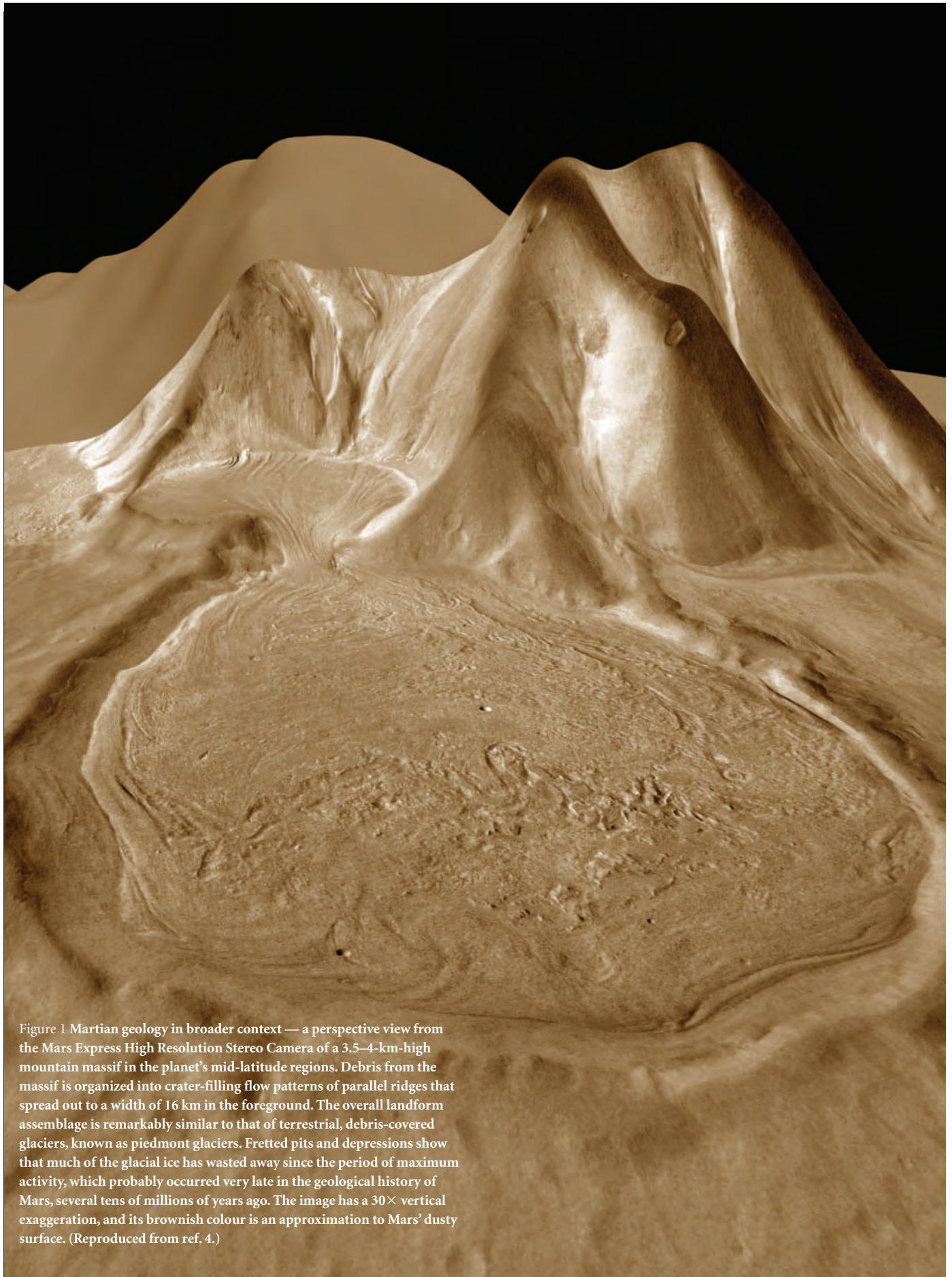


Figure 1 Martian geology in broader context — a perspective view from the Mars Express High Resolution Stereo Camera of a 3.5–4-km-high mountain massif in the planet's mid-latitude regions. Debris from the massif is organized into crater-filling flow patterns of parallel ridges that spread out to a width of 16 km in the foreground. The overall landform assemblage is remarkably similar to that of terrestrial, debris-covered glaciers, known as piedmont glaciers. Fretted pits and depressions show that much of the glacial ice has wasted away since the period of maximum activity, which probably occurred very late in the geological history of Mars, several tens of millions of years ago. The image has a 30× vertical exaggeration, and its brownish colour is an approximation to Mars' dusty surface. (Reproduced from ref. 4.)

the Olympus Mons escarpment, as recently as four million years ago.

The three papers in this issue use HRSC data to provide an overwhelming case for new thinking about recent geological activity on Mars. On page 356, Hauber *et al.*³ document that explosive eruptions, about 350 million years ago, created depressions on the flank of the volcano Hecates Tholus. As recently as five million years ago, glacial deposits were laid down on the floors of these depressions.

Head *et al.*⁴ (page 346) use HRSC data to reanalyse controversial landforms that had been recognized as possibly glacial from Viking data. Combined with data from the Mars Global Surveyor, the new images (Fig. 1) show that the landforms are indeed evidence for geologically recent and recurring glacial activity at what are now tropical and mid-latitude regions on Mars. This glaciation may be a response to recent changes in the incidence of sunlight induced by variations in obliquity of the planet's spin axis¹³. Glaciers form by the net accumulation of snow and its metamorphosis to ice during an Earth-like hydrological cycle.

Early evidence from Viking data for relatively young Mars glaciation¹⁴, plus other indications of young flowing¹⁵ and 'ponded' water activity¹⁶, led to a hypothesis in 1991, considered outrageous at the time¹⁷, that Mars had been episodically active in a hydrological sense for geologically short intervals of post-Noachian time. In the 1991 hypothesis¹⁸, martian climate change is triggered by the outburst of volcanic gases and cataclysmic flood water, with the water ponding in huge seas that subsequently evaporated.

Murray *et al.*⁵ (page 352) present striking evidence for just such an event, resulting in a frozen water body approximately the size of Earth's North Sea. The inundation was generated within the past five million years because of cataclysmic flooding that accompanied volcanic eruptions from the fracture system associated with two long and narrow depressions known as Cerberus Fossae. Flood water was probably conveyed to the sea via the Athabasca Valles channel system at peak flows of about ten million cubic metres per second⁹. This may be a smaller version of what happened when much larger channels delivered peak flows of hundreds of millions of cubic metres per second to inundations on the northern plains of Mars during much earlier periods of post-Noachian martian history¹⁸.

What happened to all the water that induced the past, short-term hydrological cycling on Mars? The evidence from HRSC for recent aqueous activity suggests that the water is still present, as ice on the ground and water deep beneath the surface. In early May, Mars Express will deploy its Mars Advanced Radar for Subsurface and Ionosphere Sounding (MARSIS), which should

be able to detect subsurface water and ice to depths of several kilometres. Evidence from the latest pictures indicates that the water will surely be there.

Victor R. Baker is in the Department of Hydrology and Water Resources, and the Lunar and Planetary Laboratory, University of Arizona, Tucson, Arizona 85721-0011, USA.
e-mail: baker@hwr.arizona.edu

1. Squyres, S. W. *et al.* *Science* **306**, 1709–1714 (2004).
2. Neukum, G., Ivanov, B. A. & Hartmann, W. K. *Space Sci. Rev.* **96**, 55–86 (2001).
3. Hauber, E. *et al.* *Nature* **434**, 356–361 (2005).
4. Head, J. W. *et al.* *Nature* **434**, 346–351 (2005).
5. Murray, J. B. *et al.* *Nature* **434**, 352–356 (2005).

6. Leighton, R. B. & Murray, B. C. *Science* **153**, 136–144 (1966).
7. Baker, V. R. *The Channels of Mars* (Univ. Texas Press, Austin, 1982).
8. Malin, M. C. & Edgett, K. S. *Science* **288**, 2330–2335 (2000).
9. Burr, D. M., Grier, J. A., McEwen, A. S. & Keszthelyi, L. P. *Icarus* **159**, 53–71 (2002).
10. Berman, D. C. & Hartmann, W. K. *Icarus* **159**, 1–17 (2002).
11. Baker, V. R. *Nature* **412**, 228–236 (2001).
12. Neukum, G. *et al.* *Nature* **432**, 971–979 (2004).
13. Head, J. W. *et al.* *Nature* **426**, 792–802 (2003).
14. Kargel, J. S. & Strom, R. S. *Geology* **20**, 3–7 (1992).
15. Gulick, V. C. & Baker, V. R. *Nature* **341**, 514–516 (1989).
16. Parker, T. J., Saunders, R. S. & Schneeberger, D. M. *Icarus* **82**, 111–145 (1989).
17. Kerr, R. A. *Science* **259**, 910–911 (1993).
18. Baker, V. R. *et al.* *Nature* **352**, 589–594 (1991).

Evolution

Deep-sea spiral fantasies

John Cooke

Pictures of strange, gelatinous deep-sea worms have intrigued zoologists, as they hinted at the solution to an evolutionary puzzle. But does the first specimen to be obtained in good condition back the theories up?

Zoologists dream of connecting up the evolutionary pathways among the ragbag of phyla that have been placed together in the super-phylum Deuterostomia¹. This super-phylum includes the chordates — back-boned animals such as ourselves — and the hemichordates ('half chordates'), which are mostly small, bottom-living marine organisms. Hemichordates arguably possess two chordate features: gill

slits and a hollow, dorsal nerve cord. But in other ways, the two main hemichordate classes — the microscopic, tube-dwelling 'pterobranchs', which feed on suspended particles, and the much larger, mud-swallowing 'acorn' worms (the enteropneusts) — differ considerably from each other in lifestyle and anatomy. A reassessment of the evolutionary histories of the pterobranchs and enteropneusts is now prompted by



Figure 1 Worm turns — the first photograph, taken in 1962, of an abyssal enteropneust worm lying on the sea bed near the eastern wall of the Kermadec Trench in the southwestern Pacific. The worm's body ends shortly past the first bend, where the faecal coil begins. (Reproduced from ref. 4.)



100 YEARS AGO

It is stated in the *Times* that the committee, presided over by Mr. Haldane, M.P., appointed to consider the allocation of the increased grant-in-aid of education of a university standard in arts and science has now finished its inquiry... the committee proposes that the sum of 45,000/ be allotted as follows: — Manchester, 6000/; University College, London, 5000/; Liverpool, 5000/; Birmingham, 4500/; Leeds, 4000/; King's College, London, 3900/; Newcastle-on-Tyne, 3000/; Nottingham, 2900/; Sheffield, 2300/; Bedford College, London, 2000/. ... The committee expresses the view that the time has come for making a new departure in the principle on which State assistance is to be given to the highest education. It is recommended that a moderate sum should be set aside for distribution by way of payment to post-graduate students from the university colleges who devote themselves for one, two, or three years to special problems; and that to ensure the money being applied most efficiently to the stimulation of individual study, ... the distribution should assume the form of a grant made directly to the student on the advice of some impartial authority. From *Nature* 16 March 1905.

50 YEARS AGO

Medical findings during the inquiry into the recent *Comet* disasters have suggested that the possibility of lung damage by impact with a water surface at the terminal velocity of fall (about 160 ft./sec.) should be investigated... Penney and Bickley suggested that the lungs of any man or any animal would be severely injured by transmission of the shock wave set up if the chest wall were flung inwards with a velocity of 20 m./sec. (66 ft./sec.) acquired in 0.5 m.sec. or less... the critical velocity of impact is 99 ft./sec. for a guinea pig, 118 ft./sec for a mouse and 94 ft./sec. for a man... Some animal experiments were devised for testing the validity of these deductions. A vertical catapult was specially constructed for projecting guinea pigs, belly first, into a large tank of water... Grave injury to the lungs, and other viscera, was found in animals projected into water at velocities greater than 104 ft./sec., both anaesthetized and dead specimens giving the same result. ... The closeness of the agreement must be regarded as largely fortuitous in view of the approximate nature of the calculations. From *Nature* 19 March 1955.

Holland *et al.*² (page 374 of this issue), who report the collection of the first good-quality specimen of a 'lophenteropneust' — the supposed 'missing link' between the two groups.

The story starts with photographs of the sea bed taken by US oceanographers in the 1960s using a camera dangled from a ship into an abyss in the southwestern Pacific³. These dramatically showed large spiral coils on the ocean bottom, and one photograph was thought to have actually caught an animal in the act of creating one of the enigmatic patterns (Fig. 1). The creature was tentatively identified as an enteropneust worm about a metre in length, with a transparent, gelatinous body 5 centimetres thick and a laterally swollen head end. It was a giant in comparison with shallow-water acorn worms and, rather than burrowing in the sea-bed ooze as the shallow-water worms do, it appeared on the surface of the mud. The characteristic spiral or sometimes looping faecal trail was proposed to be made as the worm moved forward, swallowing sediment and defecating. The evenly separated spiral or looping pattern reflects the side-to-side movement of the head as it forages for particles at the surface of the sludge. These animals turned out to be quite common and widespread denizens of the deep, as shown by photographs of spiral and looping traces at other locations on the abyssal sea floor, particularly in the Southern Hemisphere⁴ (see Fig. 3 on page 375 for more examples).

A few years later, Danish deep-sea biologists led by Henning Lemche studied these and other photographs archived at the Scripps Institution of Oceanography in San Diego, California. They came to the startling conclusion that the broadly expanded collar region at the worm's head end bore tentacles resembling the 'lophophore' used for feeding by other pre-chordate groups such as the pterobranchs⁵. The Danish researchers proposed that the worm had a pterobranch-like head end on an enteropneust body. On this basis, they tentatively designated these deep-sea worms as a new group of hemichordates that they termed the 'Lophenteropneusta', or lophophore-bearing enteropneusts. There was a degree of tacit acceptance of this intermediary body form¹, from which some people inferred that the ancestral hemichordates were worm-like. But what was needed to determine whether this worm was indeed a living link between enteropneusts and pterobranchs was the recovery of a good specimen of one of these fragile worms for detailed examination.

Holland *et al.*² filmed a lophenteropneust gliding over the northeastern Pacific sea bed at a depth of 1,901 m before collecting it using a remotely operated vehicle. These pictures and the authors' careful anatomical study reveal no evidence for any tentacle-like structure on the broad collar on the worm, which is from a new family, genus and

species. Holland *et al.* also review a large number of deep-sea photographs showing broad-collared enteropneusts (including those examined by Lemche *et al.*), and conclude that none of the creatures has tentacles. The previous misinterpretation from low-quality photographs highlights the risks of trying to construe more than is perhaps wise from sparse and imperfect data.

Other questions, such as how these spiral traces seem to appear in isolation on the muddy ooze, are becoming clear from recent work. Time-lapse photography at 4,100 m depth off California⁶ suggests that the worms swim or drift to new feeding stations, rather than burrowing to them. The photographs show an enteropneust sweeping up detrital food using its collar, and forming a spiral faecal trace, over a period of 39 hours. The worm then completely emptied its gut and floated off the sea floor. Why don't these abyssal enteropneusts burrow? Perhaps it is because detrital food is limited in amount and quality at these depths, and swimming to a new location rather than burrowing would allow a much wider foraging range.

Other enteropneust species, however, might burrow. Another acorn worm was fortuitously recovered at 2,100 m in the deep northeastern Atlantic using a box corer (specially designed equipment for taking undisturbed samples from the top of the sea floor). The worm was found beneath a mound structure surrounded by burrow openings⁷. Moreover, numbers of yet another, unidentified, enteropneust species were found living on the surface of a soft, sandy contourite sediment at 850–1,000 m on the eastern flank of the Faroe–Shetland Channel off Scotland⁸, and multi-opening burrows were found deeper in the channel. But it is unclear whether the worms are responsible for the burrows.

It seems, then, that tentacle-bearing 'lophenteropneusts' can be relegated to the realm of fantasy, and enteropneusts and pterobranchs are not as closely related as some people had supposed. Whether or not the ancestral deuterostome body plan resembled today's worm-like enteropneusts is still unclear; what is more certain is that their body form and way of life reflect their modern-day habitats. ■

John Gage is in the Dunstaffnage Marine Laboratory, Scottish Association for Marine Science, Oban PA37 1QA, UK.

e-mail: John.Gage@sams.ac.uk

1. Barnes, R. S. K. (ed.) *The Diversity of Living Organisms* (Blackwell Science, Oxford, 1998).
2. Holland, N. D. *et al.* *Nature* **434**, 374–376 (2005).
3. Bourne, D. W. & Heezen, B. C. *Science* **150**, 60–63 (1965).
4. Heezen, B. C. & Hollister, C. D. *The Face of the Deep* (Oxford Univ. Press, 1971).
5. Lemche, H., Hanssen, B., Madsen, F. J., Tendal, O. S. & Wolff, T. *Vindensk. Meddr. Dansk Naturh. Foren.* **139**, 262–336 (1976).
6. Smith, K. L., Holland, N. D. & Ruhl, H. A. *Deep-Sea Res.* (in press).
7. Mauviel, A., Juniper, S. K. & Sibuet, M. *Deep-Sea Res.* **34**, 329–335 (1987).
8. Bett, B. J. *Cont. Shelf Res.* **21**, 917–956 (2001).

Applied physics

Electrons held in a queue

Dmitri V. Averin

When electric currents are made sufficiently small, the electrons can be seen moving one by one. This is accomplished in a microelectronic circuit, providing a means of obtaining an accurate standard for current.

Electrons carry electric charge and so repel each other because of Coulomb forces. In the case of ordinary metallic conductors, for example, when we measure a current passing through a millimetre-thick copper wire, this effect is too weak to be noticed. However, in carefully designed conducting structures with nanometre-sized features, Coulomb repulsion breaks the flow of current into separate charges, making it possible to detect and control their motion. Depending on the nature of the conductors, the individual charges can be either electrons or more complex particles, such as 'Cooper pairs' of electrons in superconductors.

Phenomena of this kind have been intensively studied, and are known as single-electron tunnelling (SET)^{1,2}. They typically take place in systems of small 'tunnel' junctions — metallic islands separated by thin insulating layers through which electrons can infrequently jump or tunnel from one island to another. Coulomb repulsion plays the dominant role in these structures and makes electrons tunnel one by one, in a correlated fashion. In particular, if a small average current flows through a tunnel junction, electron-tunnelling events should be separated in time by well-defined intervals, creating current oscillations. Despite the fact that SET effects have been investigated in detail for the past 20 years, such oscillations have been measured only indirectly^{3,4}. On page 361 of this issue, Bylander *et al.*⁵ report an experiment with a one-dimensional array of tunnel junctions in which SET oscillations are observed directly — in real time — as periodic, discrete spikes of current, each associated with one electron.

Conceptually, the simplest structure that should exhibit single-electron oscillations is a one-dimensional channel of finite length filled with electrons⁶ (Fig. 1a). If the electron density is sufficiently low, Coulomb repulsion overcomes the quantum-mechanical fluctuations of the electrons and forces them to form a well-ordered queue in which each electron occupies a well-defined position relative to the other electrons. Such a periodic structure of charged particles, minimizing the total Coulomb energy, is called a Wigner crystal. The whole line of electrons can be brought into motion by injecting an electric current into the channel. If the injected current I is constant in time, electrons arrive at any given point in the channel periodically, one after another, and the

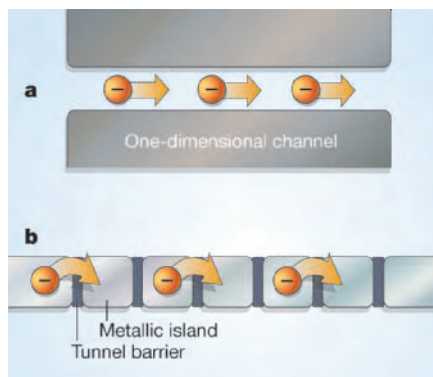


Figure 1 Lining up electrons: one-dimensional structures in which correlations between electrons are important. When a structure carries a constant average current, oscillations in the current are produced owing to the motion of the correlated electrons. **a**, A narrow channel with low electron density in which electrons form a well-ordered queue. **b**, An array of tunnel junctions used in the experiment by Bylander *et al.*⁵ — small metallic islands separated by insulating layers through which electrons can tunnel.

charge at this point oscillates with frequency $f = I/e$, where e is the charge of an electron.

Although it might seem straightforward to carry out such an experiment by measuring current passing through a sufficiently thin semiconductor wire, in practice this is difficult. When the electron concentration in a real wire is made low enough to form a Wigner crystal, impurities start to dominate the electrostatic properties of the wire, causing disorder. As a result, the correlations in electron motion are suppressed. At larger electron densities, when the effect of impurities is less important, the electrons are unable to form a Wigner crystal because of quantum-mechanical fluctuations in their positions — in that case, only weak traces of the current oscillations should remain⁷.

Bylander *et al.*⁵ used an SET device in which electrons behave qualitatively similarly to a one-dimensional Wigner crystal but without disorder and quantum-mechanical fluctuations causing serious problems. In this experiment, electrons are confined to move in a one-dimensional array of tunnel junctions (Fig. 1b). As with typical SET devices, each metallic island in the array consists of a very large number of atoms, so that the Coulomb forces between electrons propagating through the array can be described in terms of the electrostatic

capacitances of the islands. The capacitances are small enough to produce noticeable potentials when the islands are charged with individual electrons.

The change in electrostatic potential that is associated with individual electrons creates correlations between them, similar to those in a Wigner crystal. During propagation along the array, an electron raises the potential of the island it is occupying as well as that of surrounding islands, thereby preventing other electrons from tunnelling into the area. If the average current I through the array is sufficiently small, so that the typical time e/I between the tunnelling of successive electrons is much larger than the characteristic time of one electron crossing a tunnel barrier, the random nature of the tunnelling process is not important and electron order created by Coulomb repulsion is preserved in the array. Then, if the charge transferred through the array can be measured with a time resolution better than the time interval between successive electrons, as in Bylander and colleagues' experiment⁵, individual electrons can be detected arriving at the end of the array periodically in time with the period e/I .

SET oscillations, and single-charge tunnelling phenomena in general, lend themselves to new and unusual applications. One of the first areas where SET devices have already made the transition to application is thermometry⁸. The devices are similar to the arrays used by Bylander *et al.*, which at higher temperatures can operate as absolute thermometers. Such a thermometer relates temperature directly to another, better-defined quantity (in this case, voltage across the array of metallic islands) by a fundamental relation and does not require calibration.

Single-electron oscillations offer a similar possibility for electric current. By relating the current I to the frequency f of the oscillations through the fundamental constant e ($I = ef$), a well-defined, 'quantum' standard of current is obtained. Because frequency can be determined with much better precision than current, such a standard would improve the accuracy of the definition of current. The main problem with realizing such a standard with SET oscillations, or with other SET-based approaches⁹, is that a higher absolute value for the current is required. At the maximum frequency where electrons are correlated (typically in the range of 10^6 Hz), the current obtained from $I = ef$ is still too small for applications. A possible solution to this problem would be to combine many devices in parallel, thereby increasing the total current. Bylander *et al.* suggest that it should be easier to accomplish this with their device, but further work is needed to see whether using SET oscillations is indeed the most promising way to obtain a definitive current standard.

In a different form of application, Bylander and colleagues' demonstration that it is

possible to rapidly count individual electrons could be used to study the statistics of charge transfer through other SET devices, or other types of highly resistive conductors. This is particularly interesting for systems of tunnel junctions between superconducting islands, where the dynamics of individual Cooper pairs is governed by quantum mechanics. If the time resolution of measurements could be further improved, non-trivial quantum measurements could then be performed. These might demonstrate various quantum paradoxes in the dynamics of macroscopic systems, and advance current efforts to develop scalable quantum information processing devices. ■

Dmitri V. Averin is in the Department of Physics and Astronomy, Stony Brook University,

State University of New York, Stony Brook, New York 11794, USA.

e-mail: daverin@qcom.physics.sunysb.edu

1. Averin, D. V. & Likharev, K. K. in *Mesoscopic Phenomena in Solids* (eds Altshuler, B. L., Lee, P. A. & Webb, R. B.) 173–271 (Elsevier, Amsterdam, 1991).
2. Sohn, L. L., Kouwenhoven, L. P. & Schön, G. (eds) *Mesoscopic Electron Transport* (Kluwer, Dordrecht, 1997).
3. Delsing, P., Likharev, K. K., Kuzmin, L. S. & Claeson, T. *Phys. Rev. Lett.* **63**, 1861–1864 (1989).
4. Cleland, A. N., Esteve, D., Urbina, C. & Devoret, M. H. *Appl. Phys. Lett.* **61**, 2820–2822 (1992).
5. Bylander, J., Duty, T. & Delsing, P. *Nature* **434**, 361–364 (2005).
6. Averin, D. V. & Likharev, K. K. in *Nanostructures and Mesoscopic Systems* (eds Kirk, W. P. & Reed, M. A.) 283–295 (Academic, Boston, 1992).
7. Averin, D. V. & Nazarov, Yu. V. *Phys. Rev. B* **49**, 2951–2954 (1994).
8. Pekola, J., Schoelkopf, R. & Ullom, J. *Phys. Today* May, 41–47 (2004).
9. Kautz, R. L., Keller, M. W. & Martinis, J. M. *Phys. Rev. B* **62**, 15888–15902 (2000).

Cancer biology

Sense out of missense

Terry Van Dyke

The p53 protein is notorious for its involvement in many cancers. Studies in mice are helping to clarify how mutations in the human p53 protein produce a wide variety of tumours.

The gene encoding the p53 protein is mutated in more than 50% of human cancers, but quite how it is involved in tumour biology has eluded scientists for decades. Studies in *Cell* by Olive *et al.*¹ and Lang *et al.*², of genetically engineered p53 mutant proteins in mice, underscore the complexity of p53 cancer mechanisms, and provide a model for a human cancer-susceptibility syndrome.

The p53 protein regulates gene expression in response to a variety of cellular stresses, and can either stop cell growth or cause cells to die (for recent reviews, see refs 3, 4). And yet it was thought to promote cell growth until analysis of human tumours showed that of the two copies of the *p53* gene that are normally present, one is often missing and the remaining copy is frequently mutated⁵. Indeed, the p53 protein originally analysed was a mutant version. Further examination defined p53 as a tumour suppressor, a protein that inhibits cancer processes. End of story? On the contrary, just the beginning!

Although normal p53 acts to suppress tumours, p53 mutants, rather than just relieving the suppression, can actively promote tumour growth. Furthermore, surprisingly few of the *p53* mutations delete chunks of the protein; most mutations are missense mutations that result in a change in a single amino acid, leaving the rest of the protein intact. *In vitro* studies (reviewed in ref. 6) suggested two possible, not necessarily mutually exclusive, explanations for these puzzling observations. First, many p53 mutant proteins can clearly inhibit normal

p53 function. Second, and more controversially, the mutant proteins might acquire new abnormal functions. As always with *in vitro* assays, the results could be due to experimental conditions. Moreover, such studies cannot address the specific role(s) of putative gain-of-function mutants in the development of cancer, or whether the effects are specific to particular cell types. Olive *et al.*¹ and Lang *et al.*² have tackled these issues using *in vivo* approaches.

Missense mutations can arise throughout much of the p53 protein, but certain 'hot-spot' mutations appear frequently. Two such mutations, termed R270H and R172H, were examined in mice by Olive *et al.*; Lang *et al.* independently analysed the R172H mutation. The mutants represent the two general classes of p53 missense mutation: contact mutants, including R270H, in which the structure of p53 is mostly preserved but the protein's DNA-binding function is perturbed; and structural mutants, such as R172H, in which the overall p53 structure is affected. In addition, the equivalent mutations in human p53 occur in people with Li–Fraumeni syndrome, a condition in which an inherited p53 mutation predisposes the individual to a variety of early-onset cancers^{7,8}.

Both studies show clear gain-of-function, tumour-promoting activities *in vivo* and *in vitro*. Olive *et al.*¹ found that both mutations predisposed mice to a spectrum of tumours that is distinct from that found in mice lacking one (*p53*^{+/-}) or both (*p53*^{-/-}) copies of *p53*. This was true whether or not the mice had a normal copy of *p53* in

addition to the mutant one (*p53*^{+/-} or *p53*^{-/-}). Most notably, both mutations produced an increase in the frequency of carcinomas, the cancers that are derived from epithelial cells. Moreover, the carcinomas were more aggressive in mice harbouring mutant *p53* than in *p53*^{+/-} or *p53*^{-/-} mice.

In contrast, Lang *et al.*² did not observe an increase in carcinomas in their *p53*^{-/-} mice. But the carcinomas that did occur often spread quickly, unlike those in their *p53*^{+/-} counterparts. As the two studies used different mouse strains, it is possible that other genes influence the effects of the p53 mutant proteins. Such a mechanism could explain why epidemiological studies of Li–Fraumeni patients yielded conflicting conclusions about whether p53 missense mutants have gain-of-function activities. Unlike laboratory mouse strains, humans are genetically diverse, and complex genetic effects are difficult to pinpoint.

The two missense mutations produced distinct tumour patterns, including different frequencies of particular tumour types and varying properties of specific tumour types¹. This reflects varying effects of the p53 mutations on the different cell types from which the tumours arise, possibly because of the difference between contact and structural mutants, or because each mutant might have unique tumorigenic properties.

Do gain-of-function activities involve interaction with normal p53? It appears that some may and some do not. Clearly, the differences between the *p53*^{+/-} and *p53*^{-/-} mice require no interaction with normal p53; however, differences between *p53*^{+/-} and *p53*^{+/-} mice could involve an alteration of normal p53 activities, although there are effects that cannot be accounted for by simple inactivation. Furthermore, these outcomes could themselves be specific to the mechanism and/or the cell type. For example, comparing embryonic fibroblast cells derived from *p53*^{+/-} mice with those from *p53*^{-/-} mice shows both similarities and differences, depending on the factors being examined: basal levels of cell proliferation are increased in both, but certain aspects of the cell cycle and the programmed cell-death pathway vary dramatically between them¹. Moreover, during embryonic development, *p53*^{+/-} mice are similar to *p53*^{-/-} mice in response to DNA damage, but not in the interaction with MDM2, a protein that inhibits normal p53 (ref. 2).

What are the molecular mechanisms involved in gain-of-function p53 effects? Studies in mouse embryonic cells and tumour-derived cell lines implicate the interaction of mutant p53 with relatives of p53 called p63 and p73 (reviewed in ref. 9). Reducing the levels of p63 and p73 in *p53*^{-/-} embryonic cells produced similar effects on growth rate and uncontrolled growth as in embryonic cells harbouring the p53 R172H

mutant². Moreover, p53 R172H interacts with p63 and p73 in tumour-derived cells^{1,2} and inhibits activation of their target genes².

It has long been assumed that mutant p53 proteins were inherently stable, leading to higher levels of the protein. However, surprisingly, both studies found that the p53 mutants examined were minimally present in normal tissues and became stabilized only in tumours. This result indicates that a change in the stability of mutant p53 protein occurs during tumour evolution, suggesting that a key p53 regulatory network must be altered for the mutant protein to be effective. This finding is novel and provocative, and may be crucial in understanding the overall mechanism of tumorigenesis driven by missense p53.

Although the p53 mechanisms appear complex, with significant cell-type specificity and potential effects from other genes, the studies by Olive, Lang and colleagues provide compelling evidence for gain-of-function

tumorigenic activity in two common missense p53 proteins. Furthermore, this work paves the way for more detailed assessment of how this protein functions in cancer, and provides the most accurate preclinical models of Li-Fraumeni syndrome to date.

Terry Van Dyke is at the Lineberger Comprehensive Cancer Center, University of North Carolina Chapel Hill School of Medicine, Chapel Hill, North Carolina 27599, USA.

e-mail: tvdlab@med.unc.edu

1. Olive, K. P. *et al.* *Cell* **119**, 847–860 (2004).
2. Lang, G. A. *et al.* *Cell* **119**, 861–872 (2004).
3. Ryan, K. M., Phillips, A. C. & Voudsen, K. H. *Curr. Opin. Cell Biol.* **13**, 332–337 (2001).
4. Hofseth, L. J., Hussain, S. P. & Harris, C. C. *Trends Pharmacol. Sci.* **25**, 177–181 (2004).
5. Greenblatt, M. S., Bennett, W. P., Hollstein, M. & Harris, C. C. *Cancer Res.* **54**, 4855–4878 (1994).
6. Sigal, A. & Rotter, V. *Cancer Res.* **60**, 6788–6793 (2000).
7. Varley, J. M. *Hum. Mutat.* **21**, 313–320 (2003); Erratum *Hum. Mutat.* **21**, 551 (2003).
8. Evans, S. C. & Lozano, G. *Mol. Med. Today* **3**, 390–395 (1997).
9. Melino, G., Lu, X., Gasco, M., Crook, T. & Knight, R. A. *Trends Biochem. Sci.* **28**, 663–670 (2003).

complicating comparisons among regions. But Hinds *et al.* chose a scheme that is uniform among genomic regions for the vast majority of SNPs. Their data are therefore perfectly suited for identifying genomic regions with reduced levels of variability, regions that are enriched in population-specific differences, and regions that have undergone natural selection.

Hinds *et al.* find that correlation among different genetic variants — or ‘linkage disequilibrium’ — is enhanced in regions that contain genes as opposed to non-coding regions of DNA, and among SNPs that cause amino-acid changes in an encoded protein. This observation provides indirect evidence for the action of natural selection on at least some of the variants in the coding regions. Many factors have influenced human genetic variation, including demographic processes of population growth, subdivision and migration. However, some genes have also been targeted by natural selection as humans have adapted to new environments, fought pathogens⁵, and developed advanced cognitive abilities and culture⁶. If selection has occurred very recently, it leaves a distinct pattern in the genome — for example by increasing the frequency of particular variants or the degree of genetic differentiation among populations — and this pattern can be distinguished from the background pattern of variation using statistical methods. Further statistical analyses of Hinds and colleagues’ data may identify the specific genes or regions that have been targeted by natural selection in the recent evolutionary history of mankind.

Such studies are of interest not only because they may help to elucidate the molecular forces underlying human evolution, but also because regions targeted by selection may be associated with genetic diseases or disease susceptibility⁷. Disease-causing mutations are likely to be selected against, as long as the disease reduces the fertility or viability of the affected individual. However, some genetic-disease factors may also have increased in frequency in the population because they give the carriers a fitness advantage^{8–10}. A classic example is mutations in the β -globin gene, in which individuals carrying

Figure 1 Spot the difference. Variation in single nucleotide polymorphisms (SNPs) may be the underlying cause for variability in susceptibility to disease and in physical characteristics such as eye colour.

Human genomics

Disclosure of variation

Rasmus Nielsen

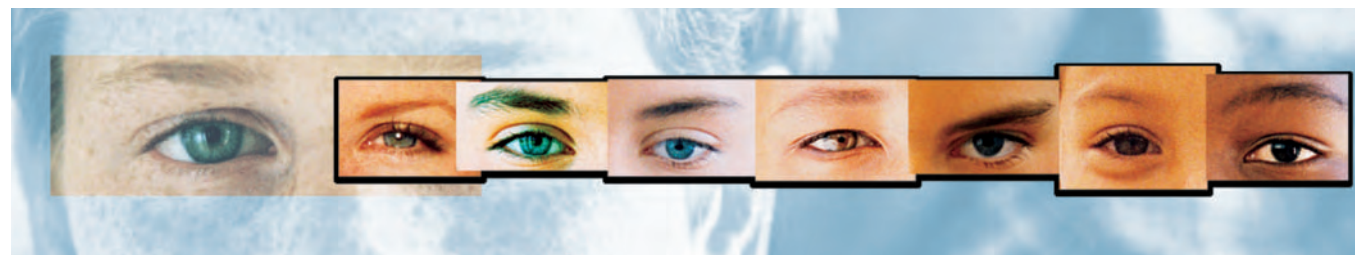
Now that the sequence of the human genome is almost complete, the human genomics community is turning its attention towards what, genetically speaking, makes people different.

In one of the most considerable studies of human genetic variation so far, scientists at Perlegen Sciences and their collaborators have quantified the variation at more than 1.5 million DNA positions in 71 individuals of European American, African American and Chinese ancestry. Their findings are described in *Science*¹. Not too long ago, genetic inferences about human ancestry were based primarily on a single gene²; now we have data on variants from millions of positions in the human genome that may help us to uncover human genetic ancestry and realize the promise of ‘personalized’ medicine.

There are roughly 10 million positions in the human genome that vary among individuals. To a great extent, the genetic causes for differences in physical and mental characteristics — including susceptibility to certain diseases — will probably be found in these ‘single nucleotide polymorphisms’ (SNPs; Fig. 1). The focus of large-scale genomics

studies is now being turned towards identifying and describing this variation, one goal being to develop new treatment options for many diseases. In the process, we are learning much about our evolutionary and demographic history, through statistical analyses of the current genetic make-up of human populations.

Hinds *et al.*¹ provide the first comprehensive study of human DNA variation in which the same SNP ascertainment scheme was used throughout the genome. An ascertainment scheme is the method by which SNPs are chosen. It effectively provides a filter that influences the patterns observed in the data³ — much like the colour of the shades in sunglasses selectively filters out certain wavelengths of light, changing the colour of the sky for the person using them. In many large-scale studies of human genomic variation, including the public HapMap project⁴, different ascertainment schemes have been used in different genomic regions,



two copies of a disease mutation develop sickle-cell anaemia, but individuals with one copy⁸ have partial resistance against malaria. It has also been proposed that the mutations that cause cystic fibrosis⁹ and Tay–Sachs¹⁰ disease might reduce susceptibility to typhoid and tuberculosis, respectively, in carriers with only one copy of the mutation. Detailed statistical and evolutionary analyses of Hinds and colleagues' data may suggest which other genomic regions, or genes, are most likely to harbour disease-associated mutations, even without the help of any data on the consequences of the mutations.

The availability of large-scale SNP data therefore brings together the disciplines of medical and evolutionary genetics. Evolutionary thinking underlies many of the common methods used for identifying associations between genetic types and observable traits or diseases using population genetic data, and has led to major advances in genetic epidemiology^{11,12}. And with the increased emphasis

on identifying the determinants of genetic diseases comes the awareness that human genomic variation can only be understood fully in light of the evolutionary forces that have shaped it. The data published by Hinds *et al.*¹ will provide a unique resource for explorations in both disciplines. ■

Rasmus Nielsen is at the Centre for Bioinformatics, University of Copenhagen, Universitetsparken 15, 2100 Kbh Ø, Copenhagen, Denmark.
e-mail: rasmus@binf.ku.dk

1. Hinds, D. A. *et al.* *Science* **307**, 1072–1079 (2005).
2. Cann, R. A., Stoneking, M. & Wilson, A. C. *Nature* **325**, 31–36 (1987).
3. Nielsen, R., Todd, M. J. & Clark, A. G. *Genetics* **168**, 2373–2382 (2004).
4. Gibbs, R. A. *et al.* *Nature* **426**, 789–796 (2003).
5. Sawyer, S. L. *et al.* *PLoS Biol.* **2**, e275 (2004).
6. Enard, W. *et al.* *Nature* **418**, 869–872 (2002).
7. Halushka, M. K. *et al.* *Nature Genet.* **22**, 239–247 (1999).
8. Friedman, M. J. & Trager, W. *Sci. Am.* **244**, 154–164 (1981).
9. Pier, G. B. *et al.* *Nature* **393**, 79–82 (1998).
10. Rotter, J. I. & Diamond, J. M. *Nature* **329**, 289–290 (1987).
11. Slatkin, M. *Genetics* **137**, 331–336 (1994).
12. Pritchard, J. K., Stephens, M., Rosenberg, N. A. & Donnelly, P. *Am. J. Hum. Genet.* **67**, 170–181 (2000).

Language

Syntax for free?

Ricard Solé

Human language is based on syntax, a complex set of rules about how words can be combined. In theory, the emergence of syntactic communication might have been a comparatively straightforward process.

Language does not fossilize — for all that it was one of the great transitions in evolution¹, the advent of language has left no obvious equivalent to fossil teeth and bones, and seems inaccessible to enquiry. But it is not hard to imagine the emergence of a set of signals to label objects, the combinatorial nature of which allowed an infinite repertoire of sentences to be constructed from a finite set of words. An essential part of the process must have been the acknowledgement of a set of rules to combine words in such a way as to make sentences meaningful. These rules are the syntax that we all easily learn as children, but students of language evolution have a tough time explaining its origins.

So, how did syntax originate? Some authors have conjectured that it resulted from some pre-adaptation of the human brain² — that the first step was the building of a rich lexicon upon a cognitive system predisposed to formulate rules able to exploit the underlying combinatorial features of word associations. According to a study by Ramon Ferrer i Cancho and colleagues³, however, a simple word–object association matrix can provide the basis for syntax almost for free.

The authors use a simple model in which signals (words) are associated with objects. Such association can be referential (such as *meat* referring to 'edible organic matter') or

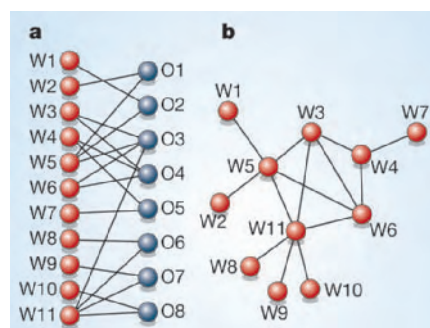


Figure 1 Building a protolanguage network. a, A bipartite set of connections can be built by linking signals (words, red) to objects (blue). Most words are specific, referring to only one or two objects, whereas a few of them have many links. b, A new network is obtained by linking words that share at least one object of reference. The resulting network has many words with few links, but some acting as hubs. Ferrer i Cancho *et al.*³ believe that syntactic rules might have emerged from such a scale-free network architecture.

non-referential (*eat* is associated with 'action of eating'). Some words display both associations: *eat*, for example, is also linked to 'edible organic matter'. A bipartite set of connections is obtained linking the two sets (Fig. 1a).

Ferrer i Cancho *et al.* show that most of the architecture of this network stems from a

universal feature common to all languages, Zipf's law. Briefly stated, Zipf's law establishes that if we take all the words in a text and order them by rank, from the most common to the rarest, the frequency (number of appearances) decays inversely with their rank⁴. Most words are rare, whereas a few (such as *the*, *of*, *and*, *to*, *I*) are very common. Common words are less specific and can be linked to many objects, whereas rare ones are more specific and have one or a few links. The number of links of a given word, not surprisingly, follows the same law.

This set of linkages does not provide any obvious clue to the origins of syntax. But it allows another network to be built that incorporates a primitive form of word–word association. The new network naturally links names and actions, because pairs such as *eat* and *meat* will be connected. This is simply done: two words will be linked if they share at least one object of reference (Fig. 1b). This is of course a very rough way of associating symbols, but, as Ferrer i Cancho *et al.* show, the architecture of the resulting network seems to be surprisingly close to many features exhibited by linguistic networks.

Previous studies⁵ have demonstrated that networks obtained by linking words exhibit seemingly universal patterns of organization, provided the words are syntactically related: two words are linked if they have been syntactically combined in a collection of sentences. Different languages share the same scale-free structure, with most words having few syntactic links and a few of them being connected to many others.

Ferrer i Cancho and colleagues' network displays the same structure as its real counterparts. Exactly the same distribution of links is found, suggesting the possibility that early 'protolanguage' might have been ready-made for the development of a full syntax. If this is so, the sometimes illogical and quirky appearance of syntactic rules might be nothing but a by-product of scale-free network architecture. The study also suggests that Zipf's law could have been a precondition for syntax and symbolic communication. Once such a condition was met, the basis for the combinatorial explosion characteristic of human language was ready for selection to shape it. The new theory will be subject to debate, but the remnants of the communicative Big Bang are evidently hiding somewhere inside modern language networks. ■

Ricard Solé is in the Complex Systems Laboratory, ICREA-Universitat Pompeu Fabra, Barcelona 08003, Spain.
e-mail: ricard.sole@upf.edu

1. Maynard-Smith, J. & Szathmáry, E. *The Major Transitions in Evolution* (Oxford Univ. Press, 1995).
2. Bickerton, D. *Language and Species* (Univ. Chicago Press, 1990).
3. Ferrer i Cancho, R., Riordan, O. & Bollobás, B. *Proc. R. Soc. Lond. B* doi:10.1098/rspb.2004.2957 (2005).
4. Zipf, G. K. *Human Behaviour and the Principle of Least Effort* (Addison-Wesley, Cambridge, Massachusetts, 1949).
5. Ferrer i Cancho, R. *et al.* *Phys. Rev. E* **69**, 051915 (2003).

Bioengineering

Joint endeavours with titanium

Proc. Natl Acad. Sci. USA

doi:10.1073/pnas.0500693102 (2005)

An increasing number of people receive artificial joints each year, and not only the elderly. But replacement joints themselves sometimes have to be replaced, one reason being that the commonly used titanium components can loosen over time.

K.-L. Paul Sung and colleagues have taken a closer look at why titanium joints become loose. They found that titanium particles produced by the wear and tear of the implants disrupt bone-building cells known as osteoblasts. Microscopic observations revealed that the particles become concentrated inside these cells, preventing them from attaching to the artificial joint. The researchers also implanted pins in rat thighs and found that the presence of titanium particles made the pins come out more easily.

The results exemplify why scientists team up with clinicians to produce better materials for human implants. Using nanotechnology, Sung and colleagues are now at work developing implant material that has three to five times higher wear resistance than current options.

Roxanne Khamisi

Chemistry

Stimulating syntheses

J. Am. Chem. Soc. doi:10.1021/ja0422007 (2005)

When it comes to biologically active molecules, nature — naturally — is the first place to look. These 'natural products' occur in microbes, plants and animals worldwide, and many are in use as treatments for a wide range of diseases. But synthesizing these molecules is rarely straightforward: they are often topologically complex, which means that chemists must find creative ways to minimize the number of steps required.

Pengfei Wang *et al.* report the first synthesis of QS-21A, the natural source of which is the bark of a South American tree. QS-21A is a potent immunostimulatory agent, used in more than 80 clinical trials (including trials of vaccines for various cancers, HIV-1 and malaria).

Like many natural products, the molecule is decorated with carbohydrates, which are often necessary to elicit the desired biological activity. It can be difficult to synthesize such polysaccharides and attach them to the molecular scaffold selectively and in high yields, and Wang *et al.* had to use a variety of methods throughout the procedure — for example, 'dehydrative glycosylation', which their

Physics

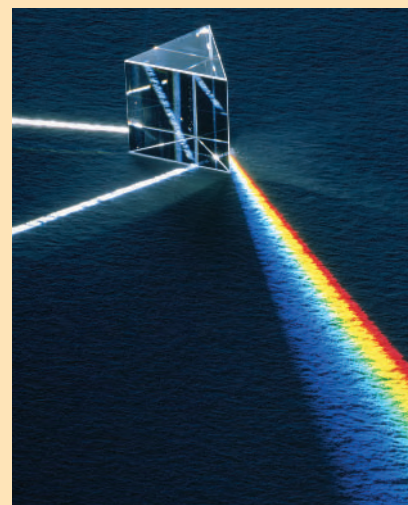
Pause for reflection

Phys. Lett. A **336**, 271–273 (2005)

Total internal reflection occurs in a glass prism when light strikes the interface between glass and air at an angle greater than a critical angle determined by the refractive index of each medium. Isaac Newton proposed that, in these circumstances, the light would be delayed in the air before re-entering the glass; the Hungarian physicist Eugene Wigner predicted a value for this delay in 1955.

Dominique Chauvat *et al.* have now measured the delay. They find that there are in fact two different 'Wigner delays', which depend on the polarization of the light.

Using femtosecond pulses of light polarized perpendicular to the plane of incidence, they found that the delay increased as the angle of incidence approached the critical angle. It reached a maximum of 28 femtoseconds before internal reflection gave way to refraction — when the beam does not return through the glass (see picture). But when the polarization of the pulse was twisted through 90° to be parallel to the plane of incidence, the delay reached 57 femtoseconds. This implies that unpolarized light experiences two



Total internal reflection (left) and refraction (right) at a prism–air interface.

separate delays on reflection, the authors argue.

Chauvat *et al.* suggest that understanding this phenomenon could be useful for probing materials that have a negative refractive index. They add that similar Wigner delays should also exist for beams of particles such as neutrons.

Mark Peplow

group has developed in several key steps for construction of the complex carbohydrate fragments.

The authors say that the synthesis of derivatives of QS-21A is under way, for use in research to explore how this remarkable natural product works *in vivo*.

Joshua Finkelstein

Neurobiology

Fragile flies

Neuron **45**, 753–764 (2005)

Fragile X syndrome is a single-gene disorder in which symptoms range from mild learning difficulties to severe mental retardation. It affects about 1 in 6,000 people, making it the most common form of inherited mental retardation. Using fruitflies to model the condition, Sean M. J. McBride, Catherine H. Choi and colleagues find that several neurological and behavioural deficits can be reversed pharmacologically.

Studies of mice lacking the causal *FMR1* gene suggest that overactive signalling from neuroreceptors called group I mGluRs underlies the condition. McBride, Choi *et al.* used fruitflies to evaluate the involvement of these receptors, as fruitflies engineered to have no *dfmr1* gene (the fly relative of *FMR1*) show defects in courtship behaviour, brain anatomy and memory that are analogous to aspects of fragile X syndrome.

The authors treated the flies with compounds that inhibit mammalian mGluRs, and found that the insects' courtship and memory, and in certain cases

their brain structures, returned to normal. The most dramatic improvements occurred when treatment began in the fly larvae and continued into adulthood.

Helen Dell

Evolutionary biology

In search of cellulases

Mol. Biol. Evol. doi:10.1093/molbev/msi107 (2005)

Many animals lack the enzymes — cellulases — able to digest the plant material cellulose; instead, they rely on symbiotic organisms in their intestines to do the job. So it is widely assumed that the ancestors of these animals also lacked cellulases, and that animals that now have the enzymes picked them up by gene transfer from elsewhere.

Angus Davison and Mark Blaxter challenge that view. They scoured genetic databases for the genes of a particular class of cellulase enzyme, 'glycosyl hydrolase family 9'. They found these genes in many more animal groups than expected, including earthworms and sea urchins. Because the genes are similar in structure in the different groups, the finding suggests that they were inherited from a common ancestor, perhaps one that existed before the divergence of plants, animals and fungi.

The implication is that the ancestors of many animals that now lack cellulases had the relevant genes but lost them at some point during evolution. It remains unclear why, but one possible reason is that digesting cellulose is not worth it given the benefits to be gained from digesting other foods.

Helen Pearson

Earthquake risk from co-seismic stress

Last year's Indonesian earthquake has increased seismic hazard in the region.

Following the massive loss of life caused by the Sumatra–Andaman earthquake in Indonesia and its tsunami, the possibility of a triggered earthquake on the contiguous Sunda trench subduction zone is a real concern. We have calculated the distributions of co-seismic stress on this zone, as well as on the neighbouring, vertical strike-slip Sumatra fault, and find an increase in stress on both structures that significantly boosts the already considerable earthquake hazard posed by them. In particular, the increased potential for a large subduction-zone event in this region, with the concomitant risk of another tsunami, makes the need for a tsunami warning system in the Indian Ocean all the more urgent.

Inspection of the aftershock distribution and evidence from recent inversions reveal that the Sumatra–Andaman earthquake of 26 December 2004 ruptured almost 0.25 million square kilometres of the Indian plate/Burma microplate subduction zone (Fig. 1), generating a tsunami. The current death toll is in the region of 300,000.

Subduction-zone earthquakes are often coupled: in the Nankai trough subduction zone to the southeast of Japan, for example, five of the seven large earthquakes on the Nankaido segment in the past 1,500 or so years were accompanied by similar events on the contiguous Tonankai/Tokai segment within five years, and three of those occurred in the same year¹. This observation is entirely consistent with stress interaction, which has been shown to explain the space–time juxtaposition of large earthquakes². The destructive Izmit earthquake (magnitude 7.4) southeast of Istanbul, for example, was triggered by stress increases of less than 2 bars that were due to earlier local events³; in turn, this triggered the Düzce earthquake (magnitude 7.1), which occurred three months later⁴.

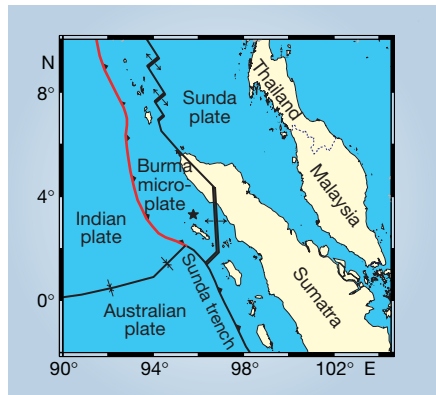


Figure 1 Plate tectonics of the Sumatra region. The red line indicates the southern extent of the surface trace of the Sumatra–Andaman earthquake rupture. Star, 2004 earthquake epicentre.

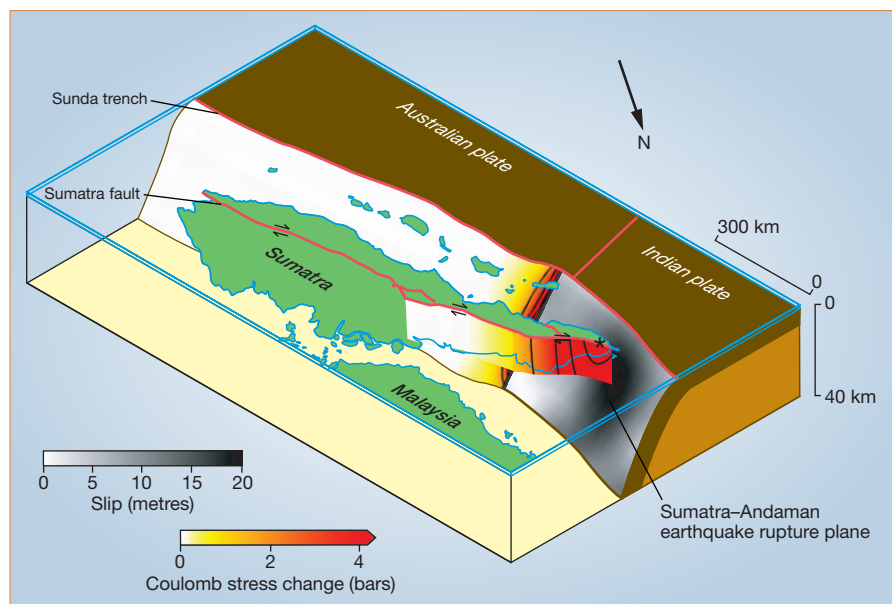


Figure 2 The Sumatran subduction zones with the overlying plates removed. Calculated three-dimensional stresses have been projected on to a diagram of the three-dimensional structural geometry and geography of the region. Grey-scale values on the rupture plane represent the amount of slip in metres experienced on the southernmost 450 km of the Sumatra–Andaman earthquake. Colour-scale values represent the co-seismic stress changes on the Sunda-trench subduction zone and the Sumatra fault. Stress contours (in black) show 2-bar intervals, starting from a maximum of 8 bars on both faults. Essential features of the calculated stresses are robust to changes in the slip distribution in recent long-period inversions, which show continuation of slip to the north for a total rupture length of about 1,200 km (ref. 10). Black asterisk indicates location of the devastated Indonesian city of Banda Aceh.

Previous work on the palaeoseismology of the Sunda trench has indicated that this area may already be advanced in the seismic cycle⁵. The northern section of the Sumatra fault has not experienced any large earthquakes for at least the past 100 years either.

Waveform-inversion studies reveal a strongly heterogeneous slip distribution for the Sumatra–Andaman earthquake, with maximum displacements being of the order of 20 m and the majority of the slip being concentrated on the southernmost 500 km or so of the rupture⁶. We used this slip distribution to calculate the stress perturbation tensor, which was then resolved on the structures of interest. Results show a stress increase of up to 5 bars in the 50 km of the Sunda trench next to the rupture zone, but they also show a strong positive loading of up to 9 bars for about 300 km on the Sumatra fault near the city of Banda Aceh (Fig. 2).

The results indicate that although a subduction-zone event in the Sunda trench has been made more likely by the Sumatra–Andaman earthquake, at present the increase in stress is localized on the north of this segment. The effect might be expected to spread further south in the months ahead as a result of viscoelastic relaxation in the lower crust, which has not been calculated here⁷. Earthquakes on the Sunda trench in

1833 and 1861 are known to have produced fatal tsunamis⁸.

The co-seismic stress perturbation on the Sumatra fault described here is significantly larger and of a greater spatial extent than the stresses that are believed to have triggered large, strike-slip earthquakes in the North Anatolian Fault Zone². Considering past activity and the observed structural complexity on the northern Sumatra fault⁹, an earthquake of magnitude 7–7.5 on this structure would seem to represent the greatest immediate threat.

John McCloskey, Suleyman S. Nalbant, Sandy Steacy

School of Environmental Sciences, University of Ulster, Coleraine, County Derry BT52 1SA, UK
e-mail: j.mccloskey@ulster.ac.uk

1. Rydelek, P. A. & Sacks, I. S. *Earth Planet. Sci. Lett.* **206**, 289–296 (2003).
2. Stein, R. S., Barka, A. A. & Dieterich, J. H. *Geophys. J. Int.* **128**, 594–604 (1997).
3. Nalbant, S. S., Hubert, A. & King, G. C. P. *J. Geophys. Res.* **103**, 24469–24486 (1998).
4. Hubert-Ferrari, A. et al. *Nature* **404**, 269–272 (2000).
5. Zachariasen, J., Sieh, K., Taylor, F. W., Edwards, R. L. & Hantoro, W. S. *J. Geophys. Res.* **104**, 895–919 (1999).
6. Ji, C. http://neic.usgs.gov/neis/bulletin/neic_slav_ff.html (2005).
7. Pollitz, F. F. & Sacks, I. S. *Bull. Seismol. Soc. Am.* **87**, 1–10 (1997).
8. Newcomb, K. R. & McCann, W. R. *J. Geophys. Res.* **92**, 421–439 (1987).
9. Genrich, J. F. et al. *J. Geophys. Res.* **105**, 28327–28341 (2000).
10. Stein, S. & Okal, E. *Nature* (in the press).

Competing financial interests: declared none.

Biomechanics

Independent evolution of running in vampire bats

Most tetrapods have retained terrestrial locomotion since it evolved in the Palaeozoic era^{1,2}, but bats have become so specialized for flight that they have almost lost the ability to manoeuvre on land at all^{3,4}. Vampire bats, which sneak up on their prey along the ground, are an important exception. Here we show that common vampire bats can also run by using a unique bounding gait, in which the forelimbs instead of the hindlimbs are recruited for force production as the wings are much more powerful than the legs. This ability to run seems to have evolved independently within the bat lineage.

Bats (Chiroptera) are the only mammals that fly, so their bodies differ from those of terrestrial mammals. As a result, most grounded bats can only shuffle awkwardly from a sprawled position⁴. However, the common vampire bat (*Desmodus rotundus*) walks forwards, sideways and backwards⁵, and initiates flight with a single vertical jump from standing⁶. Captive *D. rotundus* have also been found to 'hop' at speeds exceeding 2.0 metres per second⁵.

To determine whether this hopping behaviour constitutes a stereotyped running gait by *D. rotundus*, we tested five adult males on a treadmill inside a Plexiglas cage. The animals used a walking gait at low treadmill speeds (0.12 to 0.56 m s⁻¹) and a stereotyped running gait at high speeds (0.28 to 1.14 m s⁻¹). The walking gait was similar to the typical lateral-sequence walking gait of other tetrapods⁷; however, the run was different from any gait previously described (Fig. 1; for methods and video, see supplementary information). We classify

this novel gait as a run because it includes a notable aerial phase.

A tetrapod typically increases its speed while walking by increasing its stride frequency. At some transition speed, animals switch to a running gait that permits a further increase in speed, but at stride frequencies that are lower than would be predicted for high-speed walking^{8,9}. Our kinematic data from *D. rotundus* fit this general stride-frequency-velocity relationship. In Fig. 2, the slopes of the stride-frequency-velocity regressions, which are best fits to the walking and running data, respectively, and are shown truncated at the intersection, are significantly different (*t*-test, *P* < 0.0001, *n* = 61). These regression lines indicate that common vampire bats, like other running tetrapods, keep their stride frequencies low by walking at low speeds and running at high speeds (Fig. 2).

The walking vampire bats used stride frequencies that were comparable to those of similarly sized terrestrial mammals (mice) over the same range of speeds (Fig. 2; blue line). When running, however, the bats used lower stride frequencies than mice⁸: this could be explained by the vampire bats' long forearms, which allow longer and fewer strides to be taken during running than can be achieved by mice.

The absence of a running gait in all other bat species so far surveyed indicates that running may have been lost early in the

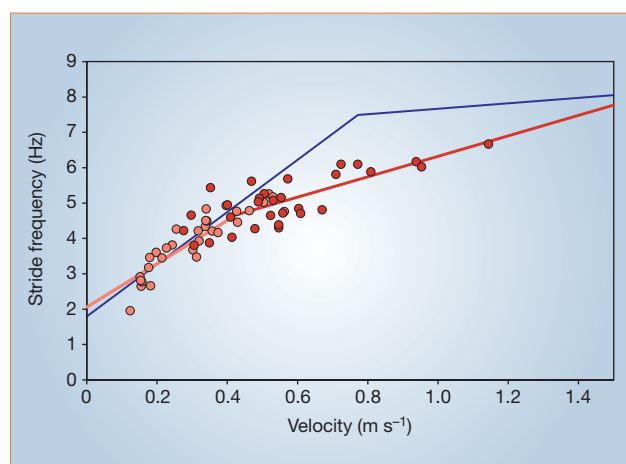


Figure 2 Stride frequency plotted against velocity for vampire bats (*Desmodus rotundus*; *n* = 5) moving on a treadmill. Pink circles, walking; red circles, running. Pink and red lines, best fits for walks and runs, respectively, truncated at their point of intersection; blue lines, best fits for walks and runs, respectively, of similarly sized (29 g) mice (data from ref. 8), shown here for comparison.

evolution of bats, evolving afresh in the vampires at a later time. We have shown that the hopping behaviour reported for *D. rotundus* in captivity⁵ is a running gait. But despite detailed knowledge of their roosting and foraging behaviour^{10,11}, the selective benefit of running for these bats in the wild is not known. Presumably, vampire bats are most likely to run when manoeuvring around prey animals while feeding, and they may have used the gait more before the introduction of domestic livestock to the Americas in the sixteenth century¹¹.

Daniel K. Riskin, John W. Hermanson

Department of Biomedical Sciences, College of Veterinary Medicine, Cornell University, Ithaca, New York 14853, USA

e-mail: dkr8@cornell.edu

- Gambaryan, P. P. *Zh. Obshch. Biol.* **63**, 426–445 (2002).
- Parchman, A. J., Reilly, S. M. & Biknevicius, A. R. *J. Exp. Biol.* **206**, 1379–1388 (2003).
- Vaughan, T. A. *Univ. Kansas Publ. Mus. Nat. Hist.* **12**, 1–153 (1959).
- Riskin, D. K., Bertram, J. E. A. & Hermanson, J. W. *J. Exp. Biol.* (in the press).
- Altenbach, J. S. *Spec. Publ. Am. Soc. Mammal.* **6**, 137 (1979).
- Schutt, W. A. Jr et al. *J. Exp. Biol.* **200**, 3003–3012 (1997).
- Hildebrand, M. in *Functional Vertebrate Morphology* (eds Hildebrand, M., Bramble, D. M., Liem, K. F. & Wake, D. B.) 38–57 (Belknap, Harvard Univ. Press, Cambridge, 1985).
- Heglund, N. & Taylor, C. *J. Exp. Biol.* **138**, 301–318 (1988).
- Taylor, C., Heglund, N. & Maloiy, G. *J. Exp. Biol.* **97**, 1–21 (1982).
- Turner, D. C. *The Vampire Bat: A Field Study in Behavior and Ecology* (Johns Hopkins Univ. Press, Baltimore, 1975).
- Greenhall, A. M. & Schmidt, U. *Natural History of Vampire Bats* (CRC, Boca Raton, 1988).

Supplementary information accompanies the paper on *Nature's* website.

Competing financial interests: declared none.

brieF communications is intended to provide a forum for short, topical reports of general scientific interest and for online technical discussion of recently published material of particular interest to non-specialist readers (**brieF communications arising**). Priority will be given to contributions that have fewer than 500 words, 10 references and one figure. Detailed guidelines are available on *Nature's* website.

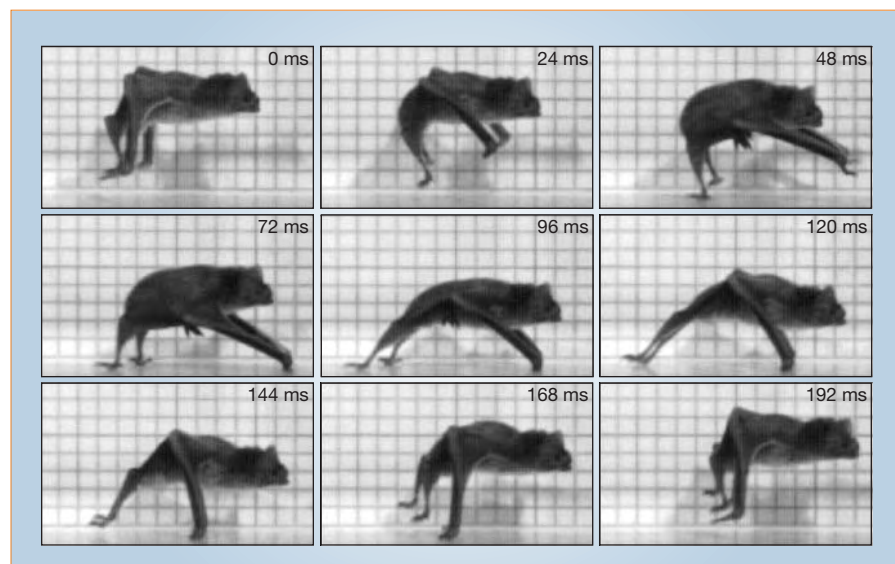


Figure 1 A vampire bat, *Desmodus rotundus*, using a running gait at 0.61 m s⁻¹ with a stride frequency of 4.71 Hz. Images are shown at 24-ms intervals; the background is a 1.0-cm² grid. For video, see supplementary information.



Cover

Richard Deacon's 2003 *Out of Order* was triggered by discussions about a major piece to celebrate 50 years of DNA; see page 308. (Courtesy of Lisson Gallery and Richard Deacon; photograph by D. Morgan.)

Editor, *Nature*:

Philip Campbell

Supplement Editors:

Alison Abbott

Adam Rutherford

Supplement Publisher:

Sarah Greaves

Production Editor:

Lucy Odling-Smee

Art Editor:

Martin Harrison

Layouts:

Vernon Adams

Picture Researcher:

Barbara Izdebska

Diagrams:

Debbie Maizels

Production:

Sue Gray

Sponsorship:

Claire Hines

Marketing:

Claire Aspinall

We cannot entirely divide our perceptual senses from our sense of reason. But in the past century in particular, the schism between art and science has been forced upon us by specialization. Science has stopped being the preserve of gentleman hobbyists or polymath intellectuals who could straddle this divide. It is hard to find today a true artist–scientist like Leonardo da Vinci, as noted for his science and engineering skills as his *Mona Lisa* and *Last Supper*. There is just too much to know.

This culture gap has, perhaps, never been so extreme. But in the past decade there has been an increasing awareness on the part of some artists of the heritage of scientists and vice versa. This supplement is intended to reflect, and place in context, some of this awareness. Many research funding agencies now provide for ‘sci–art’ projects or artists-in-residence programmes. The Institute of Contemporary Art in London has even taken a unique stance by having a scientist-in-residence. But some artists and scientists have found each other outside these formal channels, and our contributors fall within this category.

Most are from the world of arts — writers, visual artists, a composer. They describe their very individual brushes with science and explain how these encounters are reflected in some of their works.

Two are scientists who have tried to approach the question of how, neuroscientifically speaking, we perceive beauty. The science of aesthetic perception is at its very beginning and we are light years away from being able to analyse the basic unknowns of why we respond to beauty, and why we are compelled towards its creation. But it seems that we are starting to have a better idea of how to frame relevant questions.

We would like to thank the authors for their contributions to this supplement, which we hope will stimulate discussions about creativity in arts and sciences, and their possible shared roots of intuition. We are pleased to acknowledge the financial support of the SigmaTau Foundation, which contributed towards the distribution of this supplement. As always, *Nature* carries sole responsibility for editorial content. The full content, as well as other related material is available online at www.nature.com/nature/focus/arts.

Alison Abbott and Adam Rutherford



294 Fiction informed by science

A. S. Byatt, *novelist*

297 Science in literature

Simon Mawer, *novelist*

299 A tale of two loves

Alan Lightman, *novelist*

301 The artist as neuroscientist

Patrick Cavanagh, *neuroscientist*

308 From science in art to the art of science

Martin Kemp, *art historian*

310 Experimental physics, experimental art

Ken McMullen, *artist and film-maker*

312 Music, the food of neuroscience?

Robert Zatorre, *neuroscientist*

316 The evolution of sensibility

Roger Reynolds, *composer*

320 Poetry and science: greatness in little

Peter Forbes, *science writer*

Fiction informed by science

A. S. Byatt's encounters with science shape the story and characters in her four-part series of novels.

My quartet of novels begins in 1953, the year of the coronation of the new Queen Elizabeth, and the year when Francis Crick and James Watson described the structure of DNA. In 1954 I went to Cambridge, which was later agitated by the 'two cultures' row between F. R. Leavis and C. P. Snow. That was a university-political battle. Leavis held a semi-religious belief that the study of English literature was at the centre of the idea of a university. As a young woman I felt that it was only possible to study literature and try to write it if one had no such exclusive semi-religious attitude. But it wasn't Snow's polemic that interested me in science — partly because he didn't say much about science, as opposed to its 'culture'. I had had a very narrow (possibly also thorough) education. In my final years at school, languages and literature. No science at all. But I found that the writers I cared about most were in fact deeply interested in the scientific work of their time. George Eliot's *Middlemarch* not only makes reference to Andreas Vesalius, to Bichat's theory of primary tissues like webs, to optics: it weaves them into the structure of the story, the thought and the metaphorical form of the novel. In the same way Coleridge interested himself in optics, hydraulics, consciousness and growth patterns. I met John Beer, then doing his early work on Coleridge. He was interested in the way Coleridge's coiling seasnakes in *The Rime of the Ancient Mariner* might relate to Coleridge's interest in Goethe's ideas about the spiralling forms of climbing plants. We stood on the pavement and he said to me that we were in the place where DNA had just been imaged for the first time. He made a tentative connection (metaphorical) between the DNA spiral and those earlier ones.

There was passionate discussion in literary departments at that time of T. S. Eliot's theory of the "dissociation of sensibility". Eliot believed that feeling and thought, body and

mind, were a unified sensibility until the seventeenth century. After that you could no longer "feel your thought as immediately as the odour of a rose". It is hard to recapture these odd beliefs now, but at the time I found them useful for the form of a novel about the body-mind problems of a young woman interested in her own sex-versus-intellect conflict, and also in the nature of metaphor; that is, interested in the brain's excitement about making connections between disparate things.

I felt — and thought — that there should be people in the world of my novel who were interested neither in language nor in literature. One of these was the younger brother of my heroine Frederica. He is a character who is mathematically gifted and verbally inarticulate. He sees the outside world — and defends himself against its precipitous

"I met a mathematical prodigy who solved problems by imagining a garden and placing the mathematical forms in it."

errors — by mapping it geometrically as he moves in it. He sees and constructs the forms of his vision. I read a book about poetic metaphors in which the lovely story is told of the boy who always won at chess until he described how he instantly saw the best move — which was to visualize the chess board, and all the possible moves of all the pieces, in different coloured flashing lights, and choose the strongest pattern. After he had revealed his secret he collapsed in a faint. Somewhere at the same time I met the description of Francis Galton, who could visualize an imaginary slide rule and read off answers. I met a mathematical prodigy who — she said — solved problems by "imagining a garden and placing the mathematical forms in it and releasing the problem to run among the forms". Later still — many years later — I read Galton's description of the experiment in which he asked boys at public schools and undergraduates — if they understood what he was asking — to draw their own imaginary mathematical landscapes. He reproduced many of these in colour in *Inquiries into Human Faculty*.



A. S. Byatt is a writer based in London. Her novel *Possession* won the Booker Prize in 1990. Her most recent book is the fourth in the quartet described in this article, *A Whistling Woman*, published by Vintage.

One reason all this fascinated me was that it was better than any literary description of what it felt like to seize the gestalt of a work of art — remembered, or half-constructed, or unwritten but present in the mind. I was aware of the ordering operations of my own neurons, in a shadowy way, and I tried to construct, with mnemonics, conscious patterns and colours and rhythms that represented this sense of form. I am not truly synaesthetic — but I am haunted by the ghost of the possibility. And I became interested in the networks of connections that make the foundations for thinking — numbers, geometry and grammar. So my heroine, brooding about seventeenth-century metaphors in sensual language for what was beyond the sensual world, and my mathematician, were in fact struggling with the same problem.

The second novel in the quartet is called *Still Life*. It was meant to be, and is, about biological existence — sex, birth and death. I

meant to write it without metaphors and couldn't, which is how I discovered how fundamental metaphor is to thought. It is also a novel about perception — I was reading Richard Gregory and was obsessed with how we construct what we see, how we map what we see, what the brain does and what it cannot do. In *Still Life* I invented an imaginary university which set out bravely to be a real University. Like Keele University in its early days it was designed to make all-round 'Renaissance men' of its students. Keele had a compulsory first-year course where everyone learned languages and maths. That is long gone, as are the compulsory languages at Warwick University — the students demanded 'theory' instead. My metaphor for thinking is the combing of the neurons of the brain as a bird preens the hooks and eyes of a feather. I do not think 'theory' combs robust and useful neurons as maths and languages do. I invented a vice-chancellor — a Dutchman called Wijn Nobel — who is both mathematician and grammarian, like Noam Chomsky and Steven Pinker. He decides to hold a conference about the body and the mind, in which disciplines shall break their bounds, and artists and scientists shall talk and listen to each other. He is an idealist, and he is heading for the 1960s in which students believe in revolution and spontaneity, in New Age mysticism and not in scientific precision.

The third novel of the four is *Babel Tower*, and if *Still Life* was concerned with the life of the body, *Babel Tower* was designed to look at the life of the mind, especially the nature of language. I had been interested and somewhat perturbed in the 1960s by the growth of self-referring closed language-systems: sociology, psychoanalysis, hip and cool, and marxism, with their concomitant jargons. By the time I managed to write the novel, the closed languages had burgeoned into thorny thickets — semiotics, structuralism, deconstruction, lacanian psychology, feminist politics, and so on. They couldn't mostly be written into the novel; they came too late. But I had a moment of revelation when I realized that there was in fact a 'hard-wired' language that was universally human, and indeed extended beyond the human and united all the creatures of paradise or the planet. This was the four-letter code of DNA.

The third novel in the quartet, *Babel Tower*, looks at the nature of language.

Thinking about this, I discovered a solid metaphor which I embodied in the language and the narrative of my novel. I realized, one idle morning, that a snail in Latin is *helix*. And a snail's shell is in the form of a spiral. Later I discovered that there were two species of snail, *Helix hortensis* and *Helix nemoralis* (the snails of the garden and the grove), that could be fitted into both my paradise garden imagery and my realist scientific tale. By pure luck I met Steve Jones, an evolutionary

"I felt that the Fibonacci spiral was an example of a platonic order — a sense that an invisible mathematical order informed all our physical accidental world."

biologist at University College London, on a science radio programme (we were actually talking about Marcel Proust and the concept of time in physics). I discovered that Steve was the world expert on what had (unfortunately for my verbal web) been renamed *Cepaea hortensis* and *Cepaea nemoralis*. He had been studying the genetics of the external spiral of colours on the shells of the snails — work which the discovery of methods to extract DNA had rendered redundant. Novelists invent facts because of intellectual needs. I later asked Steve if he could see any connection at all between snails and work on neurons in the brain, on memory: he said that snails had giant neurons which made them peculiarly apt for this kind of experiment. I had an imagined woman scientist whom I needed to move from snail genetics to neuroscience. Curiosity is a profound drive in both novelists and scientists. I took great pleasure in learning about snails.

The other spiral that obsessed me was the Fibonacci spiral. It seemed to my non-mathematical brain a thing that could be made as a word game: take a number, add it

to itself, the next number is the sum of the previous two, and so on. But this spiral informed (to use an old seventeenth-century word for shaping from within, like the soul in the body) all sorts of natural phenomena, from climbing plants to the sprouting of twigs round stems, from snails to pine cones and sunflowers. I discovered that Alan Turing had been obsessed by explaining this and had not had the computers to do it. I met John Maynard Smith at a Darwin seminar at the London School of Economics and Political Science and he sent me a paper by two French scientists in which they work out the maths and the mechanics of growth in biological Fibonacci spirals. I cannot really understand it, but I do try. I felt that the Fibonacci spiral was an example of a platonic order — a sense that an invisible mathematical order informed all our physical accidental world. My fearful mathematician at the end of the third novel moves from studying the computer as a brain to studying this spiral. This is for him a kind of paradisaal completeness.

Writing *Babel Tower* meant going into all sorts of theories of learning and language. I was intuitively attracted to Chomsky's idea that language — a universal grammar — was 'hard-wired' in the brain. I remembered learning to read, I remembered learning French, and 'finding' rather than grimly docketing the shapes of the words and sentences. A book that has meant a great deal to me is Massimo Piattelli-Palmarini's edition of the debate in 1975 between Jean Piaget and Chomsky, at which all sorts of other thinkers from all sorts of disciplines were present, including Jean-Pierre Changeux who makes maps and metaphors of mental activity that move me as poems do. This



Inspired by spirals: an invisible mathematical order seems to inform our physical world.

“The quartet changed in the writing from a backward glance at the power of Shakespeare’s and Milton’s English and England, to a form excited by the mystery of scientific discovery.”

he was the *cauda pavonis*, the peacock’s tail of multi-coloured light before the single white light of the opus, the philosopher’s stone. (Lysgaard is a common Danish name meaning garden of light, a paradisaical reference.) I was able to make him discuss Darwin’s disgust at the idea of the useless peacock tail, which I had not known about when I named him. His research subject is also dramatically connected with his personal life.

The two visiting speakers, who hate each other, are Hodder Pinsky and Theobald Eichenbaum. They are named, following the Norse mythological pattern of the novel, after Hodur, the blind god, and Balder, the ‘dying god’ whom Hodur kills accidentally with a mistletoe spear. In the scientific structure Pinsky was a cross between Pinker and Chomsky (although his work is more related to Ulrich Neisser’s *Cognitive Psychology*), and Eichenbaum is part of the ethologist world of Konrad Lorenz and Niko Tinbergen. Eichenbaum is suspect because he worked as a scientist in Nazi Germany and even published papers using Nazi language.

The paper that Pinsky gives is entitled “Metaphors for the matter of mind”. He explains that he is “interested in a science of mind that dealt with things that were only approximately objects of language at all”. He talks about the helpful, and simultaneously misleading, nature of the image of the atom as a miniature solar system. He talks about mechanical images of the mind, about how it is both helpful and unhelpful to think of the mind by using an image of a computer. Late in my writing I read Jean-Pierre Dupuy’s *The Mechanization of the Mind* — a 1994 study of the Macy conferences on cybernetics held between 1946 and 1953 on the relation between artificial intelligence, ‘information’ (a scientific concept that draws on the seventeenth-century neo-platonic idea of the forms that ‘inform’ bodies and minds) and the brain. Cybernetics is itself a metaphor derived from the Greek for the tiller of a ship. Someone at those conferences — I think it was John von Neumann — was interested in why we enjoy puns, and suggested that it was because of the double, connecting/conflicting stimulation of a single synapse. I think that may be why we love metaphor, as well as using it. It excites our neuronal networks. I hope it is clear that Pinsky’s imaginary paper on the metaphors of mind touches on the recurring mysteries my novel looks at. (Maybe embodies.)

conference — although too late for the novels, which end in 1970 — was the kind of ideal university event my grammarian vice-chancellor wanted his body-mind conference to be. Something that became an essential part of my own thought (and of the form of my work) was the idea of models of both the physical world and the activity of the brain-mind (which is of course part of the physical world) as ‘order from noise’. And this order can be made into metaphor as the order of the shape of a flame as it burns or of a crystal as it grows. (Genes are aperiodic crystals.) Later I read in Changeux a description of the neuronal network as a combination of steady crystals of memory and ‘smoke’ of long-distance connections along the synapses. I think of these things in terms of John Donne’s angels: “So in a voice, so in a shapeless flame / Angels affect us oft and worshipped be”. But the neuronal network is real. I don’t believe in angels. I even joined a government committee on the teaching of English language (grammar) in order to understand these things at the learning level, and to put such a committee into the novels.

I ended up writing *A Whistling Woman* in the late 1990s — a novel I had expected to write in the 1970s, and this meant researching the 1960s as though they were history. I was greatly struck by the hostile attitude of the counter culture to science. I once went on a reading tour with a Barrow poet who would rise up in schools and exhort schoolchildren to pay no attention to Einstein and all that — everything they needed was in the *Prophetic Books* of William Blake. I invented such a poet and put him in the novel and wrote a poem

against science (and in favour of naked innocent childhood) in Blake’s metre.

*The metal men in coats of white
In shuttered rooms with shuttered eyes
Make metal death with metal claws
Block out the sunshine from the skies.*

*The children dance in forests free
They smell the sunshine and the rain
They dance and sing the roots and flowers
Weave magic circles whole again.*

And so on. I think it is a good encapsulation of something that was in the air.

I always knew that Professor Wijn Nobel’s body-mind conference was going to be broken up by flower children, anarchists, prophets of spontaneity. I invented an anti-university which was encamped around the ‘real’ one and taught ‘alternative’ systems such as astrology. My novel has concurrent astrological patterns, Christian patterns, scientific patterns and ordinary stories of people struggling with what feels like freedom, if limited freedom. I had huge trouble inventing the solidity of the conference itself. Here, I will briefly mention only three of the speakers.

The resident geneticist, Luk Lysgaard-Peacock gives a paper on the problem that was actually interesting Maynard Smith at the time of my conference — the profligacy of nature in constructing the male of any species. Surely, it was argued, either parthenogenesis or budding would be a more certain and economical way of passing on genes.

Lysgaard-Peacock was named originally for the alchemical thread of my patterns —

Eichenbaum's paper is never given, for the bacchic flood of demonstrators pulls him down from the podium. But I did invent his offending Nazified piece. It is a fictive essay called "Helder und Herde" (Heroes and Herds) based on Galton's chapter in *Inquiries into Human Faculty* on "Gregarious and slavish instincts". Galton studied the behaviour of African cattle and distinguished between the adventurous and curious and the purely crowd-hugging ones. Galton uses the herds as an argument for eugenics — we need to breed more independent and responsible minds. I made my Eichenbaum an expert on herd behaviour, (shoaling, flocking) — and then described his 1940s paper on these scientific facts as written in language designed to appease the National Socialists' belief in eugenics and worthless slaves. He is pulled down, as I said, by a mob, and accuses them of being "the nastiest kind of herd — tame rats in a small enclosure".

At the end of Pinsky's paper, Frederica reflects that "though she had understood what he had said, which was lucid and interesting, she was profoundly ignorant, blackly, thickly ignorant, of what he was talking about. She knew the words, neuron, synapse, dendrite, and she liked them because she could do their etymology. But the human world — including maybe some of her own forebears — had invented microscopes and telescopes, had dissected tissues and identified cells, and if it all vanished tomorrow she would not know where to start, though she might be able to write down quite a lot of *Paradise Lost* by heart (whatever her heart was, and however it worked)".

The quartet changed in the writing from a backward glance at the power of Shakespeare's and Milton's English and England, to a form excited by the mystery of scientific discovery. My world has been changed by all the scientific writers who have made their understanding approximately available to me, in plain English and working metaphors. Sometimes I talk at sci-art gatherings at which Lewis Wolpert is present. He always argues that we non-scientists do not know what we are talking about. I do not dispute this — it is true, and it is important to know it. For my part, I wish scientists were less starry-eyed about what they call 'originality' and — even more dubious — 'creativity' in the arts. But that is another essay. ■

Science in literature

Simon Mawer crosses the divide to explore how scientists and novelists alike grapple with an uncertain world.

When people discover that I have been, for much of my working life, a teacher, they evince little surprise. A teacher of literature, they assume. It's when I confess that I am a biology teacher that they are taken aback. A novelist who is a biologist is surely a contradiction in terms. Scientists are logic and facts; writers are imagination and fantasy; and between these opposite poles lies a profound gulf. Even today it is considered unusual for someone to cross the divide. Whether such an expectation originated with C. P. Snow's famous 1959 Rede lecture *The Two Cultures and the Scientific Revolution* I am not sure, but certainly his phrase has lodged in the collective mind almost as insistently as his other coining, the "corridors of power". Richard Dawkins would, one imagines, pronounce the "two cultures" a successful meme.

As a teenager with literary ambitions, but studying biology, this divide worried me. I couldn't see why a scientist should not have recourse to wit and imagination, or a writer to careful logic and reason. The fact that Snow himself was a physicist gave me little comfort — his writing seemed exactly what one might expect from the science side of his own perceived gulf: all facts and little fantasy. But elsewhere there were stirrings of hope, flutterings of possibility. As an undergraduate I heard the Nobel laureate Niko Tinbergen start a lecture course on animal behaviour with the words, "Some people try to extrapolate from our studies to human behaviour but if you wish to learn about the behaviour of man don't ask the ethologist; turn rather to the great writers. Read Dostoevsky, read Tolstoy." I paraphrase, but the sense is right.



Simon Mawer is a biology teacher in Rome. He is the author of seven novels, including the 1997 *Mendel's Dwarf*. Mawer won the 2003 Boardman Tasker Award for mountain literature for his latest novel, *The Fall*. He is currently working on his first non-fiction book — on Mendel and his legacy, for the Field Museum in Chicago.

C. WARDE-JONES

In the event it was another Russian who caught my imagination and my feeling for nature. I was studying entomology at the time, learning about the genetics of mimicry in *Papilio dardanus*, the African swallowtail, and somehow I came across this:

"The mysteries of mimicry had a special attraction for me. Its phenomena showed an artistic perfection usually associated with man-wrought things...When a certain moth resembles a certain wasp in shape and colour, it also walks and moves its antennae in a waspish, unmothlike manner. When a butterfly has to look like a leaf, not only are all the details of a leaf beautifully rendered but markings mimicking grub-bored holes are generously thrown in... I discovered in nature the nonutilitarian delights that I sought in art. Both were a form of magic, both were a game of intricate enchantment and deception"¹.

The author of these words (will a scientist reading this already have turned to the references to check?) was a research fellow at the Museum of Comparative Zoology at Harvard University. He also worked for the American Museum of Natural History and the Cornell University Museum of Entomology. As a taxonomist he named more than 20 genera, species and subspecies within the Lepidoptera and published 22 scientific papers. And yet his literary prose, dancing and dazzling and deceiving much like one of his beloved butterflies, gave birth to some of the greatest novels of the twentieth century. I refer, of course, to that most remarkable of writers in the English (and, presumably, the Russian) language, Vladimir Nabokov.

The assiduous checker of references may spot that I have carefully excised from the quotation above Nabokov's denunciation of darwinian selection. His anti-darwinian views (only slightly less notorious than his anti-freudian ones) were, I suspect, more of a protest than a postulate. He was a quantum physicist of the literary world ("I confess, I do not believe in time," he wrote) and was surely reacting against the pedestrian nature of natural selection; for there is something pedestrian about Darwin's theory. Where, after all, is the thrill in the argument with which the evolutionist greets the sceptic who doubts that you and I could have evolved by an incremental process of favoured mistakes from, say, sea squirts? "Give it enough time," he insists, "just a few hundred million years." Well, yes, I agree, but wearily. There's not much fun in deserts of vast eternity, is there? Nabokov understood, and perhaps this is what drew him to literature, exactly what many a scientist has grasped — that imagination can circumvent both dull experiment and dull exposition, and deliver ideas in a flash. For this reason — the supremacy of imagination — there has always been a cross-fertilization between science and literature. You may find it in the borrowing of 'quark' from *Finnegans Wake* ("Three quarks for Muster Mark"), or in the research interest of a major character in a famous 1960s novel:

"Ken Whitman's field of special competence, after his early interest in echinoid metabolism, was photosynthesis; his doctoral thesis had concerned the 7-carbon

sugar sedoheptulose, which occupies a momentary place within the immense chain of reactions whereby the five-sixths of the triosphosphate pool that does not form starch is returned to ribulose-5-phosphate. The process was elegant, and few men under forty were more at home than Ken upon the gigantic ladder, forged by light, that carbon dioxide descends to become carbohydrate..."².

It is there in the image of Schrödinger's cat — a purely literary conceit — stalking through the modern imagination with a smile that suggests it alone knows whether it is alive or dead. It lies in the pulse of fear that runs through one of Ian McEwan's characters when he hears the creak of a floorboard behind him:

"Nerve terminals buried deep in the tissue of the heart secrete their noradrenalin, and the heart lurches into accelerated pumping. More oxygen, more glucose, more energy, quicker thinking, stronger limbs. It's a system so ancient, developed so far back along the branchings of our mammalian and pre-mammalian past that its operations never penetrate into higher consciousness. There wouldn't be time anyway, and it wouldn't be efficient. We only get the effects. That shot to the heart appears to occur simultaneously with the perception of threat; even as the visual or auditory cortex is sorting and resolving into awareness what fell upon eye or ear, those potent droplets are falling"³.

Yet, perhaps it is only recently that some writers have made the scientific process the very focus of their work. This is what Carl Djerassi, author of several plays and novels as well as the oral contraceptive pill, would call 'science-in-fiction', to distinguish it from that tired and often tiresome genre 'sci-fi'. It is important not to be misled by the use of the word fiction here. Fiction does not stand as the antithesis to fact. Good fiction points towards truth, which is, after all, only where the scientist is trying to go. Thus Niels Bohr opens his front door to discover that Werner Heisenberg "stands on the doorstep blinking in the sudden flood of light from the house. Until this instant his thoughts have been everywhere and nowhere, like unobserved particles, through all the slits in the diffraction grating simul-

taneously. Now they have to be observed and specified"⁴.

And from this famous 1941 visit of one nuclear physicist to another, Michael Frayn goes on to explore the dilemmas of discovery in his play *Copenhagen*. I attempted something similar in my own novel *Mendel's Dwarf*, trying, like Frayn, to get to grips with the moral and ethical issues that we face today:

"The molecule has the shape of a twisted ladder," Benedict (the dwarf of the title) tells the woman who is soon to be his lover. "A Jacob's ladder, if you like, but a Jacob's ladder that goes both ways; we may use it to attempt to ascend to the throne of God... but we can also use it to descend into the pit. So beware."

'And which way are you planning to go, Dr Lambert?' Jean asks as she flips the errant lock of hair behind her ear"⁵.

Later Benedict sits at his microscope examining the embryos that he has fathered using Jean's eggs:

"He has eight embryos in eight little tubes. Four of the embryos are proto-Benedicts, proto-dwarfs; the other four are, for want of a better word, normal. How should he choose?"

How indeed? If the proper study of mankind is man, then we are bound to treat such issues through the medium of literature. But there is a great deal more to science than mere ethics: there is the whole of that imaginative construct that is the scientific vision of nature. From a textbook you get the familiar dusty story about Gregor Mendel and his peas, as though it was all a matter of logic and reason. Yet his work was one of the greatest achievements of the human intellect and imagination. It was a lifework and an obsession, a proposal that began to tease apart the very fabric of nature just as Bohr and others pulled apart the fabric of matter. Such things deserve our attention:

"Throughout each spring and summer from 1854 to 1871 the man spent hours and hours tending his plants, pollinating, scoring, labelling, harvesting, drying, putting seeds away for the next year, puzzling and pondering, counting and tallying, recording his results in leather-bound books, explaining to anyone who would listen what was going on, feeling his way into one of the greatest secrets of the natural world... You don't display obsession, not true obsession. You learn to hide it. You



A tale of two loves

The arts and sciences provide complementary ways of looking at the world, argues Alan Lightman.

recognize the indifference or incomprehension that creeps into the eyes of the listener. You learn the art of self-deprecation, the art of crypsis, the art of blending, mouse-like into the background. But beneath your bland and neutral exterior, you create confections of fantasy”⁵.

Perhaps through the medium of this passage the reader might live, for a few moments, some of those 16 years of obsession. Perhaps he might see the short, dumpy man walking in the monastery garden and thinking over his garden peas and what they might mean. This is not to miss the point of the science: it is the point. Because, just as an artistic creation lives in the mind of its creator, so too does a scientific idea. We labour under the illusion that discoveries and ideas lie somewhere out there in nature — but in truth the science is in the discoverer’s head. The inherited anlage was in Mendel’s mind, the uncertain particle was in Heisenberg’s, the Universe is in Stephen Hawking’s. That is what makes them so remarkable. There is little difference between this and artistic vision. Yes, you’ve got to do the experiments but that is not the essence of it. The essence is the idea and the enquiry. “What if?” is the question posed in both literature and science. What if the Thane of Cawdor were to have an ambitious wife and a fatal flaw in his disposition? What if a young Raskolnikov were to attempt the ultimate intellectual violence, the justification of murder? What if a grown man were to fall in love with a pubescent girl? Or... what if a gravitational field were to create a curvature of the space-time continuum? What if electrons were to be both wave and particle? What if God really were to play dice?

Darwin has brought us out of heaven and down to earth and surely it is no surprise that uncertainty — “that final core of uncertainty at the heart of things”⁴ — permeates modern literature and modern science alike. The unreliable particle and the unreliable narrator are two sides of the same weirdly spinning coin. And just as scientists employ thought experiments to focus their ideas, so a work of literature is a thought experiment about this uncertain human condition. ■

In childhood, I wrote dozens of poems. I expressed in verse my questions about death, my loneliness, my admiration for a plum-coloured sky and my unrequited love for 14-year-old girls. Reading, listening, even thinking, I was mesmerized by the sounds and the movement of words. Words could be sharp or smooth, cool, silvery, prickly to touch, blaring like a trumpet call, fluid, pitter-pattered in rhythm. And, by magic, words could create emotions and scenes. When my grandfather died, I buried my grief in writing a poem, which I showed to my grandmother a month later. She cradled my face with her veined hands and said, “It’s beautiful,” and then began weeping all over again. How could marks on a white sheet of paper contain such power and force?

Between poems, I did scientific experiments. These I conducted in the cramped little laboratory I had built out of a storage closet in my house. In my homemade alchemist’s den, I hoarded resistors and capacitors, coils of wire of various thicknesses and grades, batteries, switches, photoelectric cells, magnets, dangerous chemicals, test tubes and Petri dishes, Bunsen burners. I loved to find out how things worked.

When my experiments went awry, I could always find certain fulfillment in mathematics. When my maths teachers assigned homework, I relished the job. I would save my maths problems for last, right before bedtime — like bites of chocolate cake awaiting me after a long and dutiful meal of history and Latin. Then I would devour my cake. In algebra, I loved the abstractions, letting ‘x’s and ‘y’s stand for the number of nickels in a jar or the height of a building in the distance.

Alan Lightman is a physicist and adjunct professor of humanities at the Massachusetts Institute of Technology in Boston. He is the author of four novels, including *Einstein’s Dreams* (1993), which has been adapted for numerous theatrical and musical productions. He has published two collections of essays and six non-fiction books about physics and astronomy. His latest book, *The Discoveries*, about the great scientific discoveries of the twentieth century, will be published by Pantheon Books in November 2005.

I loved solving a set of connected equations, one logical step after another. I loved the shining purity of mathematics, the logic, the precision. I loved the certainty. With mathematics, you were guaranteed an answer, as clean and crisp as a new \$20 bill. And when you had found that answer, you were right, unquestionably right. The area of a circle is πr^2 . Period.

Mathematics and science contrasted strongly with the ambiguities and contradictions of people. The world of people had no certainty or logic. People confused me. My Aunt Jean continued to drive recklessly and

1. Nabokov, V. *Speak, Memory* (Weidenfeld & Nicolson, London, 1967).
2. Updike, J. *Couples* (Alfred A. Knopf, New York, 1968).
3. McEwan, I. *Enduring Love* (Jonathan Cape, London, 2001).
4. Frayn, M. *Copenhagen* (Methuen, London, 1998).
5. Mawer, S. *Mendel’s Dwarf* (Doubleday, London, 1997).

at great speed, even though everyone told her she would kill herself. My Uncle Edwin asked me to do a mathematical calculation that would help him run the family business with more efficiency, but when I showed him the result he brushed it aside with disdain. Blanche, the dear woman who worked for our family, deserted her husband after he abused her and then talked about him with affection for years. How does one make sense out of such actions and words?

A long time later, after I became a novelist, I realized that the ambiguities and complexities of the human psyche are what give fiction, and perhaps all art, its power. A good novel gets under our skin, provokes us and haunts us long after the first reading, precisely because we never fully understand the characters. We sweep through the narrative over and over again, searching for meaning. Compelling characters must retain a certain mystery and unfathomable depth, even for the author. Once we have seen to the bottom of their hearts, the novel is dead for us.

There are questions with answers and questions without. Scientists work on questions with answers. Although science is constantly revising itself in response to new ideas and data, at any moment each scientist is working on what is called a 'well-posed problem' — that is, a problem of such a kind and stated with such clarity that it is certain to have a definite answer. That answer may take ten years to find, or a hundred, but an answer exists. By contrast, for artists the question is often more interesting than the answer, and often an answer doesn't exist. How does one answer a question such as "What is love?" or "Would we be happier if we lived to be 1,000 years old?" One of my favourite passages from Rilke's *Letters to a Young Poet* is this: "We should try to love the questions themselves, like locked rooms and like books that are written in a very foreign tongue."

Science has much to offer the arts. When Salman Rushdie spoke to an audience at the Massachusetts Institute of Technology in late 1993, he said: "Many of us writers of my generation have felt that in many ways the cutting edge of the new is to be found in the sciences." Science has always been a source of new ideas, and artists thrive on ideas. Nicolaus Copernicus, Darwin, Einstein, James Watson

Describing the indescribable: string theorists use the language of artists to portray the fundamental nature of the Universe. (Computer artwork representing superstrings.)

and Francis Crick have all changed our view of the world and our place in it.

Then, there is the portrayal of the scientist. By now, it is well known that the picture of the scientist as the eccentric personality without human feeling, pursuing truth by the numbers, wearing sterile gloves at all times, is false. But the particular way that a person trained in logical thinking must negotiate his or her way through the illogical world of human passions — that is a subject worthy of art.

The arts and humanities, in turn, offer the sciences an essential store of other ideas, images, metaphors and language. These connections are often subtle. For his highly non-intuitive postulates of relativity, Einstein partly credited the philosopher David Hume's notion that the truths of nature can not always be arrived at by experiment but

sometimes must originate in the mind. The great atomic physicist Niels Bohr compared the invisible nucleus of an atom to an oscillating drop of liquid. Modern string theorists describe the hypothesized smallest constituents of matter and energy, which will probably never be seen by the most powerful instruments, as 'vibrating strings' that 'stretch' and 'break' and 'merge'. Such images, metaphors and vocabulary arise both from direct sensual experience and from the language of artists and humanists who portray that experience. Scientists, in turn, must use the same language to describe their extreme worlds, far beyond sensual experience, because no alternative exists except for mathematical equations.

It seems to me that the most important gift the sciences and the arts have to offer each other is a recognition and synthesis of their different approaches to thinking, their different ways of being in the world. When these differences come together, often uneasily, we witness the full complexity and the mystery, and ultimately the grandeur, of being human. What makes Michael Frayn's *Copenhagen* so powerful, in part, is the unspoken contrast between the neat world of atomic physics, where one can solve mathematical equations for wavefunctions and the neutron mean free paths, versus the world of politics, war and ethical dilemmas. And in the film *A Beautiful Mind*, the precision of John Nash's mathematical mind forms a disturbing counterpoint to the confusion in his hallucinatory illusions. When I was writing *Einstein's Dreams*, I resisted the urge to make each dream world, with its own theory of time, logically consistent. My Einstein, the most celebrated envoy of rational thought, was also a dreamer, a person of deep loneliness, struck with the frailties and urgencies of the human heart — indeed a scientist who might have failed to achieve what he did without his sometimes irrational personal commitments and passions.

Somehow, we human beings have a wondrous capacity for being both rational and irrational, detached and passionate, deliberate and spontaneous, craving of certainty and uncertainty, seeking questions with answers and questions without. We are a splattering of contradictions. In my own case, I have always felt these juxtapositions as a creative tension necessary for my work, a continual rumbling in my gut, an unsettled joy. ■

The artist as neuroscientist

Artistic licence taps into the simplified physics used by our brain to recognize everyday scenes, says Patrick Cavanagh.

Although we rarely confuse a painting for the scene it presents, we are often taken in by the vividness of the lighting and the three-dimensional (3D) layout it captures. This is not surprising for a photorealistic painting, but even very abstract paintings can convey a striking sense of space and light, despite remarkable deviations from realism.

The rules of physics that apply in a real scene are optional in a painting; they can be obeyed or ignored at the discretion of the artist to further the painting's intended effect. Some deviations, such as Picasso's skewed faces or the wildly coloured shadows in the works of Matisse and other Impressionists of the Fauvist school, are meant to be noticed as part of the style and message of the painting. There is, however, an 'alternative physics' operating in many paintings that few of us ever notice but which is just as improbable. These transgressions of standard physics — impossible shadows, colours, reflections or contours — often pass unnoticed by the viewer and do not interfere with the viewer's understanding of the scene. This is what makes them discoveries of neuroscience. Because we do not notice them, they reveal that our visual brain uses a simpler, reduced physics to understand the world. Artists use this alternative physics because these particular deviations from true physics do not matter to the viewer: the artist can take shortcuts, presenting cues more economically, and arranging surfaces and lights to suit the message of the piece rather than the requirements of the physical world.

In discovering these shortcuts artists act as research neuroscientists, and there is a great deal to be learned from tracking down their discoveries. The goal is not to expose the

Figure 1 By 1467, artists such as Fra Carnevale had mastered consistent perspective but not consistent lighting. The people in the foreground cast deep shadows but those on the plaza above and to the left do not. The alcove on the right is brightly lit but the only opening in its left wall is a small door. The shadows on the right wall of the alcove rise mysteriously upwards. These severe inconsistencies are not evident or jarring to the human viewer. (Detail of *The Birth of the Virgin* by Fra Carnevale.)

'slip-ups' of the masters, entertaining as that might be, but to understand the human brain. Art in this sense is a type of found science — science we can do simply by looking.

To count as a 'discovery' in this art-based neuroscience, deviations from standard physics must be mostly invisible to the human eye in casual viewing. A painting that, despite physical impossibilities in the depiction, gives an unhindered sense of the space and objects within it, says something about our brain. For example, a shadow that looks like a convincing shadow, even though its shape does not match the object that cast it, suggests the physics of light and shadow used by our visual brain is simpler than true physics^{1,2}.

This simplified internal physics employed by our visual brain is used not just to appreciate paintings, but to enable our rapid and

efficient perception of the real world. Real shadows are subject to an extensive set of constraints, but few of these seem to be checked by our vision; that is why an artist can use an unrealistic representation with such great impact. It is important to note that the simplified rules of physics that interest us (and the artists' shortcuts that exploit them), are not based on the ever-changing conventions of artistic representation, as they hold for monkeys and infants^{3,4}, both quite immune to the conventions of art. These simplified rules are grounded instead in the physiology of the visual brain.

Darkness alone required

Cast shadows have appeared on and off in Western art from the early classical Greek⁵ and Roman paintings and mosaics to the beginning of the modern era⁶. In contrast, with the exception of a single drawing, cast shadows did not appear in Eastern art until modern times⁷. Artists take many liberties when depicting shadows, using the wrong colour or shape, without disturbing the apparent light, space or form of the depicted scene. These physical impossibilities that slip by unnoticed (Fig. 1) are important for understanding vision. They reveal that the visual brain recognizes shadows using only a small subset of the criteria that constrain real

Figure 2 Signorelli takes great liberty with shadows, but goes too far here in making the guard's shadow cross over the satyr's shadow as if it were paint. Although shadows can lie in the wrong direction and have the wrong shape, they cannot look opaque and still appear as shadows. (Detail from *The Assumption of the Virgin with Saints Michael and Benedict* by Luca Signorelli.)

shadows. Unsurprisingly, one criterion used unfailingly by artists is that the shadows must be darker than their immediate surroundings. This finding has been confirmed by perceptual experiments¹ that examine the recovery of shapes defined by shadows. Such experiments have also shown, as have artists many times over, that few if any other deviations from realism affect the recovery of shape from shadows. Exceptions to this broad tolerance can be found in

paintings where shadows fail to be convincing. Specifically, shadows should not appear to have volume or substance of their own (Fig. 2), a criterion that has yet to be examined scientifically.

Scientific studies of the perception of shadows, and shape from shadows¹ have supported other discoveries made by painters. In the two-tone images of Fig. 3, the shadows below the nose, eyebrows and chin define the depth of the face. When the shadows violate

the rules required by the visual system, the face is no longer seen as such a strong 3D structure. These studies show, for example, that the shadows must be darker, but do not have to be of physically possible colours. Studies of lighting direction support painters' intuition that inconsistent direction of lighting is not readily noticed. The cubes in Fig. 4 are all lit from one direction, with one exception. Subjects take a long time to pick out the oddly lit cube⁸.

Figure 3 The dark red areas of the two-tone image of a man's face on the left include both regions of dark shadow and dark pigment (eyebrows, hair, moustache). These areas, in appropriate lighting, should be darker than the surrounding green areas. (If this is not the case, move to a location with fluorescent or natural lighting.) In the version on the right, the same red areas are now brighter than the green surround. Shadows have to be darker to support the recovery of object shape from shadow cues so the face on the right is much less 3D. But notice that the shadows, as long as they are darker, do not have to be the right colour (the ambient red light seen in the shadows should also fall in the green areas, making them yellower than they are)¹.



Surrogate boundaries

Much of our earliest recorded art takes the form of line drawings and, remarkably, the elements of this type of representation have remained unchanged. But given that lines do not divide objects from their backgrounds in the real world⁹, why do line drawings work?

The effectiveness of line drawings is not simply attributable to learned convention, passed on through culture. Infants¹⁰, stone-age tribesmen¹¹ and even monkeys¹² are capable of interpreting line drawings as we do. So what do lines represent to the brain? Artists do not just trace the brightness discontinuities in an image. Conventional line drawings do not include the outlines of cast shadows or pigment contours; rather, they trace out the contours that characterize shape (Fig. 5). Artists have discovered which key contours must be perceived by the visual brain for the viewer to identify the essential structure of an object. By studying the nature of lines used in line drawings, scientists may eventually gain access to this natural knowledge base.

Seeing through paint

It is not easy to draw or paint a material that is barely visible and through which background patterns are only slightly altered. Artists do this by making a reasonable version of the background surface appear through the transparent surface. This superposition involves crossing the contours of

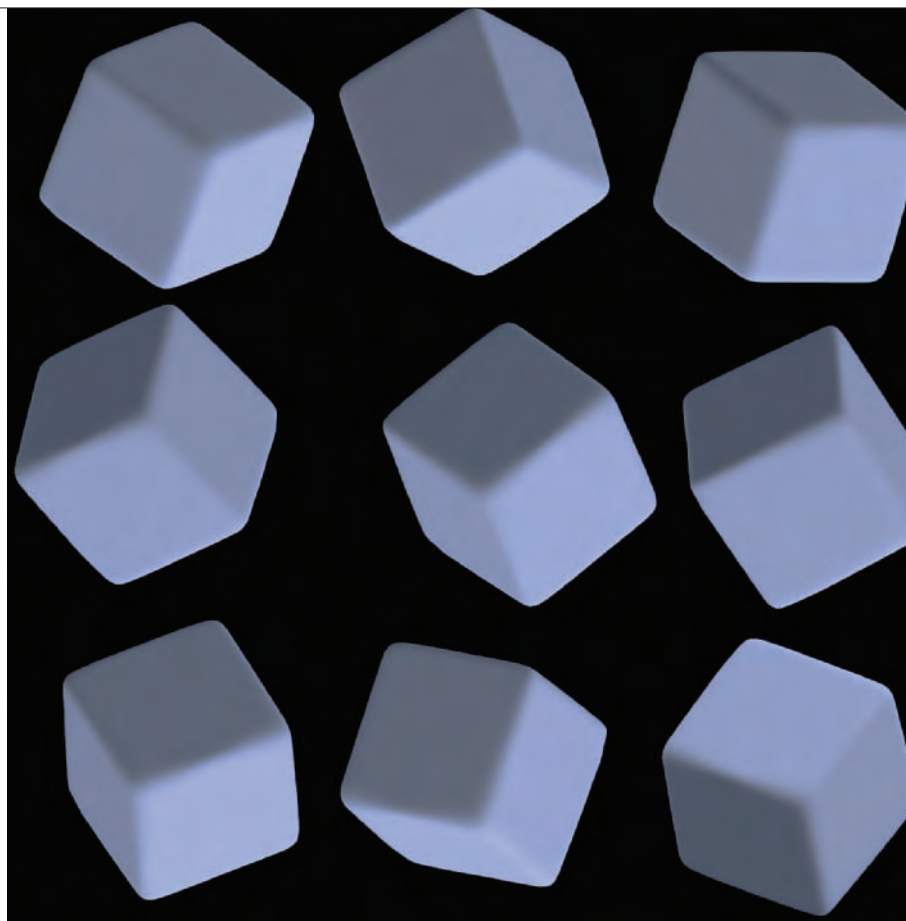


Figure 4 An array of cubes all lit from the same direction except one. Subjects take an average of eight seconds to find the odd item (bottom right here) and make many errors (30%), suggesting that inconsistent lighting is not readily noticed. (Reproduced from ref. 8.)

the transparent object with the contours in the background. For example, in Fig. 6 (overleaf), the front rim of the glass crosses the back waterline, and in Fig. 7 (overleaf), the bottom hem of the sheer cotton garment crosses the outline of the legs that are visible both above and below the hem. Experiments by Metelli have shown how these crossings or 'X-junctions' are critical cues for the successful depiction of transparency¹³. When the

X-junctions are misaligned, the impression of transparency is lost (Fig. 8, overleaf).

Although the X-junctions must be present to successfully convey transparency, other properties of the transparent material are not critical. For example, in paintings of water and glass, gross deviations from the optics of refraction (Fig. 6) are rarely noticed by the viewer, indicating again that the visual brain only computes a small set of the possible

Figure 5 Lines are used to convey the outer contours of the horses in a very similar way in these drawings, one from 15,000 BC (left: *Chinese Horse*, paleolithic cave painting at Lascaux, France) and the other from AD 1300 (right: Jen Jen-fa, detail from *The Lean Horse and the Fat Horse*, Peking Museum, China).

Figure 6 No optical distortion of the lemon in the water is shown here and yet the glass and the water appear convincingly transparent. (*Implement Blue* by Margaret Preston, 1927; oil on canvas on paperboard, 42.5 x 43 cm; gift of the artist 1960 collection; Art Gallery of New South Wales.)

physical properties of a transparent material in assessing whether or not a surface is transparent.

Filling in the gaps

Many paintings only hint at the elements in a scene and depend on the viewer's memories to construct meaningful images from the fragments. Impressionism and Cubism in particular rely on this memory-based reconstruction to complete scenes from partial representations. These paintings demonstrate the minimal skeletons of visual forms that are capable of evoking remembered images (Fig. 9).

No need for 3D

Flat paintings are so commonplace that we seldom ask why flat representations work so well. If we really experienced the world as 3D, an image seen in a flat picture would distort jarringly when we moved in front of it. But it does not do so as long as it is flat. A folded picture, in contrast, distorts as we move around it (Fig. 10). Our ability to interpret representations that are less than 3D indicates that we do not experience the visual world as truly 3D (refs 14–16), and has allowed flat pictures (and movies) to dominate our visual environment as an economical and convenient substitute for 3D representations. This tolerance of flat representations is found in all cultures¹⁰, infants³, and in other species⁴ so it cannot result from learning a convention of representation. Imagine how different our culture would be if we could not make sense of flat representations. Visual art would all be 3D: there

Figure 7 (left) Egyptian artists were the first to depict transparency. They needed to show the elegance of the fine transparent cottons worn by the wealthy (Pharaoh Sethi I on the right). The transparency of the cotton tunic is captured through overlapping contours and contrasts.

Figure 8 (far left) When a transparent surface covers a contour in the object behind it, the contour of the transparent surface and the underlying contour cross to form an X-junction. Two of these X-junctions are seen in the top panel. As Metelli showed¹³, if the contours are displaced to eliminate the X-junction, as on the bottom panel, the same patches of light and dark look more opaque than transparent.



would be no paintings or movies. Our pockets would be bulging with little statuettes of loved ones rather than photographs.

Less is more

Impressionists used minimal detail in their paintings, yet their pieces evoke a strong sense of place and mood. A photograph of an equivalent scene might be unexceptional, but the inaccurate splashes of colour and hints of contour are often moving despite, or perhaps because of, discrepancies from a realistic portrayal (Fig. 11, overleaf).

Why is this style so effective? Recent neuroscience studies of the connection between vision and the centres of emotion suggest a possible reason. Brain imaging¹⁷ of subjects presented with faces expressing fear show that the amygdala (a centre of emotion) responds strongly to a blurry version of the faces. In contrast, areas responsible for conscious face recognition respond weakly to blurry faces and best to faces presented in sharp detail (Fig. 12, overleaf). Impressionist works may connect more directly to emotional centres than to conscious image-recognition areas because the unrealistic patchwork of brush

Figure 9 Two dancers are made up of some of the isolated swatches of colour. The arrangements are sufficiently similar to familiar human shapes to trigger the integration of the marks as legs, arms, heads and bodies of single figures. Before the development of brain imaging, similarly disconnected images²³ were used by neurologists to identify brain injury to the parietal lobe. (*The Yellow Dancers* by Gino Severini, circa 1911–12; oil on canvas, 45.7 × 61 × 2.3 cm.)

strokes and mottled colouring distract conscious vision (Fig. 11).

Depicting reflection

Mirrors have been depicted in art since Greek and Roman times but, inevitably, artists commit fascinating errors when representing what is reflected by the mirror¹⁸. Having encountered reflections in mirrors throughout our lives, we might assume that

we understand how they work — what objects should be visible in the mirror given our position, the angle of the mirror and the location of other objects around us. Real mirrors never test this knowledge because they are always correct. But painters do test our knowledge of mirrors and reflection, and reveal that basically we have none. In a painting, almost any reflection will do, with only a few limits. Artists can depict people looking

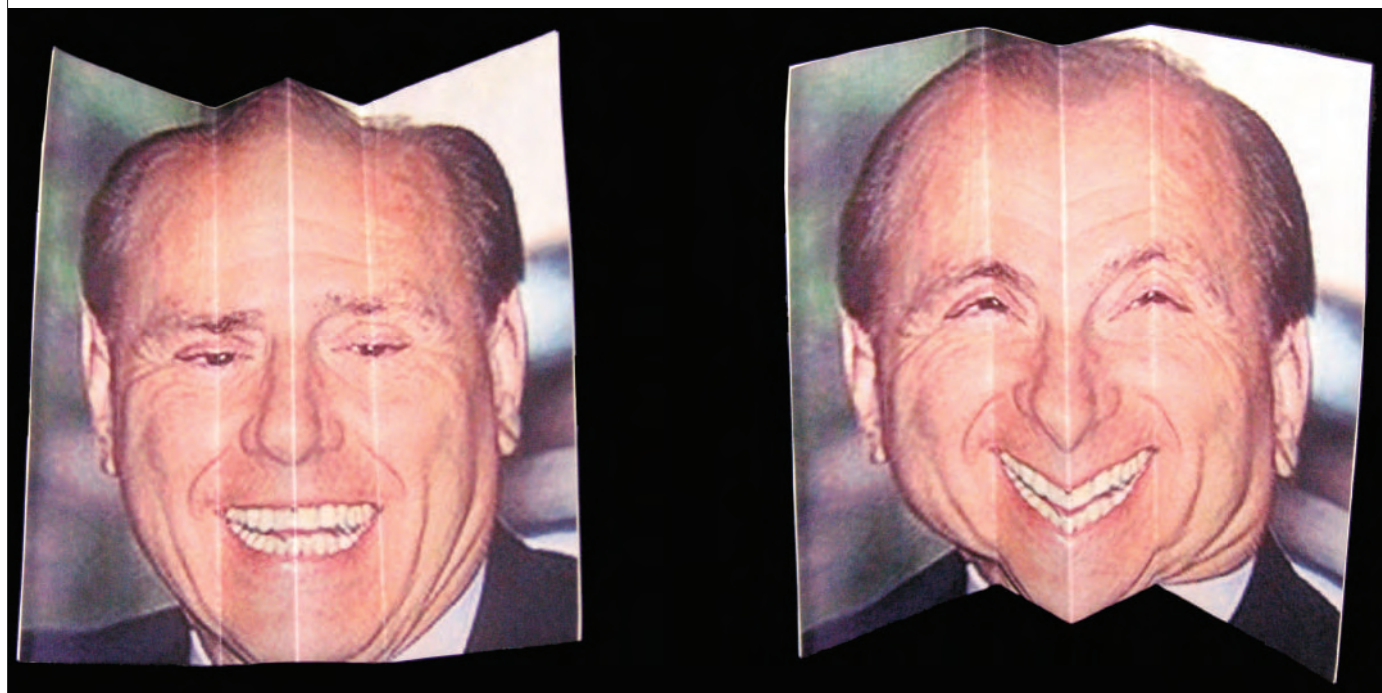


Figure 10 When a flat picture is viewed from different angles, the 3D scene can still be perceived without jarring distortions. In contrast, when a folded picture of a face is tilted, striking changes of expression are seen. (Reproduced from ref. 16.)

Figure 11 The blurry global shapes and colours may convey emotional content directly to emotional centres of the brain while the irrelevant fine detail typical of Impressionist pieces distracts conscious perception. (Pierre Renoir, 1875; oil on canvas, 55.1 × 65.9 cm; Potter Palmer Collection.)



Figure 12 Vuilleumier *et al.*¹⁷ found that the blurry, fearful face on the right activated the amygdala more than the sharply detailed or unfiltered versions. This suggests that low spatial frequencies (gross detail) provide the amygdala, an important emotional centre of the brain, with coarse, but rapid, fear-related information. The face-recognition areas of the ventral visual cortex showed less activation in response to the blurry versions than to the sharp or unfiltered images. Slower conscious analysis may therefore rely on the high spatial frequencies for face identification. Earlier studies²⁵ showed that the amygdala can respond to the fearful images even in a brief, masked presentation that subjects do not report seeing. The right amygdala has even been shown to respond to emotional facial expressions in a patient with no primary visual cortex and no conscious visual experience²¹.

at their reflections and reveal both the person and the reflection, often when this is geometrically impossible (Fig. 13).

Recent vision research also demonstrates that people have little or no awareness of where reflections ought to be¹⁹, or even of what they should look like²⁰. Subjects in one experiment had to indicate on a drawing where, when walking into a room, they would first see the reflection of a particular object in a mirror. Even physics students were unable to do this with any accuracy. Experiments also show, as painters have known for centuries, that the pattern of reflection on a surface doesn't have to match the actual scene around it for it to appear as a reflection²¹ (Fig. 14). The pattern only needs to match the average properties of natural scenes²⁰ and curve in concert with the implied curvature of the shiny surface²².

The neuroscience of art

Paintings and drawings are a 40,000-year record of experiments in visual neuroscience, exploring how depth and structure can best be conveyed in an artificial medium. Artists are driven by a desire for impact and economy: thousands of years of trial and error have revealed effective techniques that bend the laws of physics without penalty. We can look at their work to find a naive

Figure 13 Try to determine where the woman is looking. Could she be looking at her own reflection or is that physically impossible?

physics that uncovers deep and ancient insights into the workings of our brain. Discrepancies between the real world and the world depicted by artists reveal as much about the brain within us as the artist reveals about the world around us. ■

Patrick Cavanagh is in the Vision Sciences Laboratory, Department of Psychology, Harvard University, 33 Kirkland Street, Cambridge, Massachusetts 02138, USA.

1. Cavanagh, P., Leclerc, Y. G. J. *Exp. Psychol. Hum. Percept. Perform.* **15**, 3–27 (1989).
2. Casati, R. *The Shadow Club* (Little Brown, New York, 2004).
3. Yonas, A., Cleaves, W. T. & Pettersen, L. *Science* **200**, 77–79 (1978).
4. Bovet, D. & Vauclair, J. *Behav. Brain Res.* **109**, 143–65 (2000).
5. Bruno, V. J. *Form and Color in Greek Painting* (W. W. Norton &

- Company, New York, 1977).
6. Hills, P. *The Light of Early Italian Painting* (Yale Univ. Press, New Haven, 1987).
7. Gombrich, E. H. *Shadows* (Natl Gallery Pub., London, 1995).
8. Ostrovsky, Y., Sinha, P. & Cavanagh, P. *Perception* (in the press).
9. Kennedy, J. M. & Silver, J. *Perception* **3**, 313–322 (1974).
10. Yonas, A. & Arterberry, M. E. *Perception* **23**, 1427–1435 (1994).
11. Kennedy, J. M. & Ross, A. S. *Perception* **4**, 391–406 (1975).
12. Itakura, S. J. *Gen. Psychol.* **121**, 189–197 (1994).
13. Metelli, F. *Sci. Am.* **230**, 90–98 (1974).
14. Busey, T. A., Brady, N. P. & Cutting, J. E. *Percept. Psychophys.* **48**, 1–11 (1990).
15. Koenderink, J. J. & van Doorn, A. J. in *Looking into Pictures* (eds Hecht, H., Schwartz, R. & Atherton, M.) 239–300 (MIT, Cambridge, Massachusetts, 2003).
16. Cavanagh, P., von Grünau, M. & Zimmerman, L. *IEEE Proc. 2nd Intl Symp. 3D Data Processing, Visualization and Transmission* 35–41 (2004).
17. Vuilleumier, P., Armony, J. L., Driver, J. & Dolan, R. J. *Nature*

Neurosci. **6**, 624–631 (2003).

18. Miller, J. *On Reflection* (Yale Univ. Press, New Haven, 1998).
19. Croucher, C. J., Bertamini, M. & Hecht, H. J. *Exp. Psychol. Hum. Percept. Perform.* **28**, 546–562 (2002).
20. Fleming, R. W., Dror, R. O. & Adelson, E. H. *J. Vis.* **3**, 347–68 (2003).
21. Pegna, A. J., Khateb, A., Lazeyras, F. & Seghier, M. L. *Nature Neurosci.* **8**, 24–25 (2005).
22. Savarese, S., Fei-Fei, L. & Perona, P. *ACM Intl Conf. Proc. Series: Proc. 1st Symp. on applied perception in graphics and visualization* 115–118 (2004).
23. Wasserstein, J., Zappulla, R., Rosen, J., Gerstman, L. & Rock, D. *Brain Cog.* **6**, 1–14 (1987).
24. Morris, J. S., Ohman, A. & Dolan, R. J. *Proc. Natl Acad. Sci. USA* **96**, 1680–1685 (1999).

Acknowledgments: I thank M. Bernson, E. Besancon, M. Carrio, A. Dietrich, H. Farid, L. Gay, A. Kiely, A. Lonyai, C. Pemberton, D. Wang and R. M. Shapley for their contributions to this work.

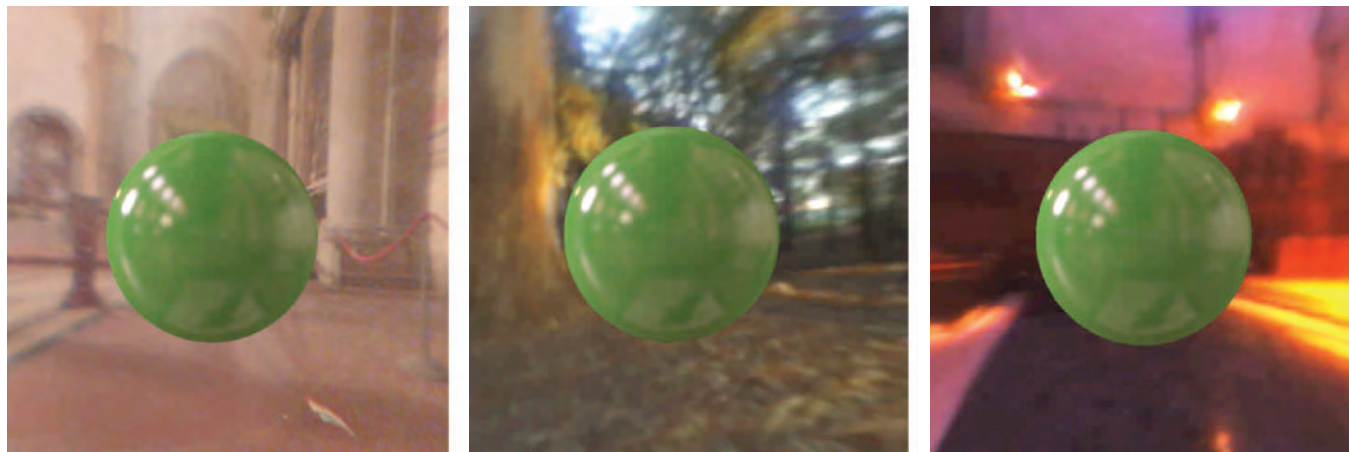


Figure 14 These green billiard balls are all shown to reflect the scene that surrounds the ball on the left²⁰. In the central and right examples, the reflections no longer correspond to the new surrounds but subjects perceive the balls to be as glossy as the one on the left. (Reproduced from ref. 20.)

From science in art to the art of science

Shared intuitions about the natural world drive the pursuits of artists and scientists, says Martin Kemp.

To generalize about the relationship between art and science is not so much hazardous as impossible. Neither science nor art are homogeneous categories. Just as science ranges from the observed complexities of environmental biology to the unseeable dimensions of theoretical physics, so art extends from the figurative representation of nature to the elusive abstractions of conceptual art. And even if we take just one area of science, say molecular biology, its expression in art may vary from straightforward iconographical reference (for instance, by allusion to the iconic double helix) to the use of molecular processes to form an unpredictable image.

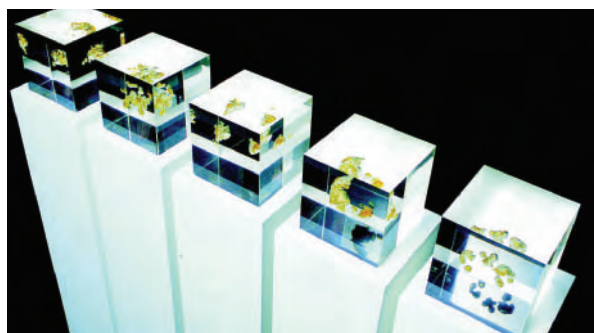
What is clear is that we serve any enquiry into art and science badly if our criterion is superficially the influence of science on art, or the influence of art on science. Deeper realms of enquiry concern complex dialogues centred on issues of cognition, perception, intuition, mental and physical structures, the communicative and social action of images, and the role of what we call the aesthetic as a shared instinct across the arts and sciences. There are as many potential relationships as there are artists who undertake profoundly imaginative scrutiny of science, and scientists who knowingly exploit their creative instincts in seeking what is 'true and beautiful'.

Historically, the most straightforward relationships have been iconographical and illustrational. The Renaissance revolution in naturalism allowed both the convincing illustration of scientific specimens and procedures, and the production of portraits of scientists in various media. No science was left untouched by the new modes of illustration, whether the descriptive science of anatomy or the three-dimensional geometries of cosmology. The stock portraits of scientists presented with suitable attributes will be familiar to anyone acquainted with older scientific institutions. As with portraits of all professionals, the image of the scientist in action only became unexceptional in the nineteenth century. This type of portrayal

obstinately persists, not least in photography, but there are now encouraging signs of artists taking a creative look at the issue of portraiture. Marc Quinn's 'portrait' of Sir John Sulston, using a culture from Sulston's own DNA, is perhaps the most striking case in point¹.

In the traditional media of sculpture, painting and drawing, artists have recently invented a mode of representation that involves neither illustration nor iconographical reference. Richard Deacon's *Out of Order* was triggered by discussions about a major piece to celebrate 50 years of DNA, but came to assume its own independent life as the artist began to think about the structures and processes that lay at the heart of the origins of life. The huge resulting complex of ribbons wheeling in space — seen on the cover of this supplement — expresses what the artist sensed to be the essence of big organic molecules without illustrating them. It is, in effect, RDNA (Richard Deacon Nucleic Acid).

Some aspect of structure is involved in all deeper dialogues between the arts and sciences. I have signalled this shared quality in my phrase 'structural intuitions', which refers to the resonances between our mental structures and patterns of organization — dynamic and static — in nature². Comparable instincts operate at all scales in the scrutiny of our natural and technological worlds, and the cosmos. The Korean artist, Nam June Paik, has reached out into cosmic realms of infinity through the endless rebounds of lasers on basic geometrical figures³. His modelling of light in space has clear affinities with physicists' and astronomers' imaginative visualizations of space-time. But the end product of Paik's installations is designed to induce a sense of awe through a suggestive extension of what can be rendered visible, rather than provide a model that is available to testing.



In *The Five Senses* by Annie Cattrell, transitory neuronal activity is caught within glassy cubes.

In quieter realms, the traditional ordering of painted and sculpted spaces through optical geometries and pythagorean harmonics, explored by Renaissance artists such as Piero della Francesca, continues to present new possibilities. Alex Colville, the Canadian painter, shows how ideas closely related to those of the Renaissance masters can exercise



Dancing with the elements: beautifully engineered components pivot around rotating axes in Susumu Shingu's *Locus of Rain*.

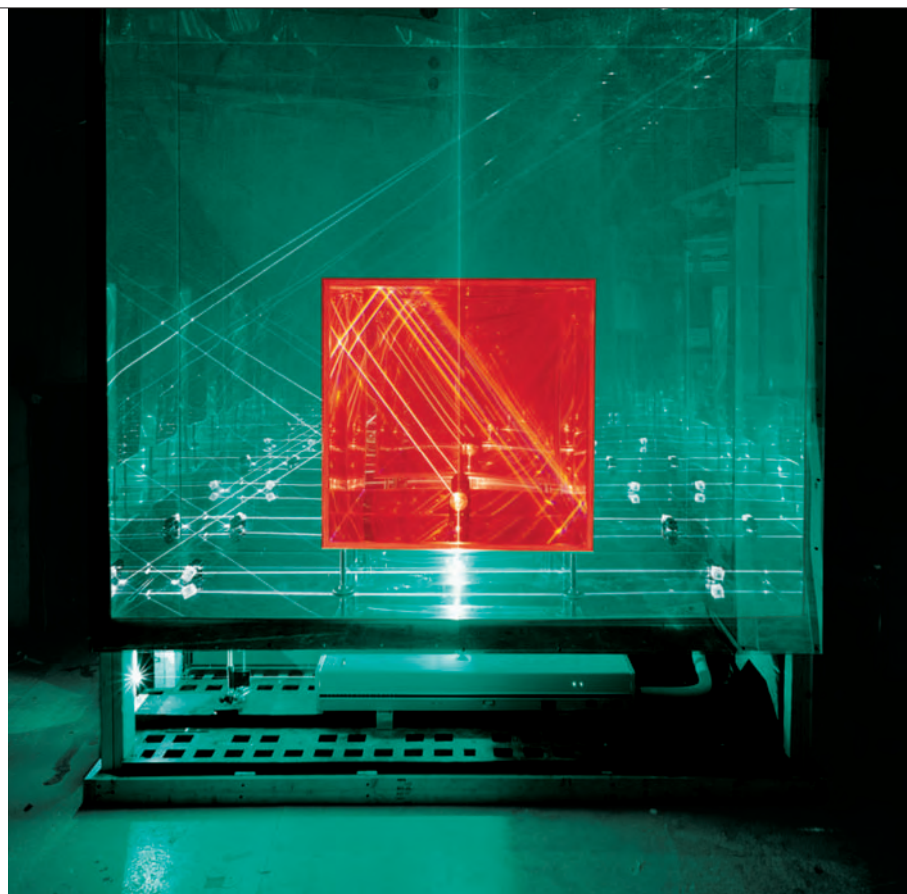
an undiminished magic, both formally and psychologically⁴. In a different physical mode, the kinetic sculptures of the Hungarian Attila Csörgő assemble the geometrical solids beloved of Johannes Kepler from apparently incoherent medleys of rods⁵. Both bring beguiling order from unsettling chaos in ways that have an enduring history.

The structures intuited by artists and scientists also involve the patterns of process in nature. In the past, painters have frozen the self-organizing motions of phenomena, such as turbulent water, into linear arabesques. Leonardo da Vinci is the great master of these effects, but he is only the foremost among many.

One of the characteristics of twentieth-century art was to embed the temporal nature of the process dynamically into the work — either so that the ‘finished’ piece was the unpredictable result of dynamic parameters set up by the artist, or such that we watch the work transforming in front of our eyes as a kinetic experience (for examples, see ref. 6).

The use of process to reveal forms of order shows obvious affinities with the mathematics of complexity. The monumental sculptures of the Japanese artist Susumu Shingu typically exploit a series of beautifully engineered components that pivot around axes, which are themselves in rotational motion⁷. When the sculptures are subject to natural forces, such as the flow of wind or fall of water, they perform fluid ballets of infinitely variable choreography within the universe of cycles and epicycles prescribed by their construction. Shingu’s sculpture outside the old Fiat factory at Lingotto in Turin relies on successively filled and erratically emptying trumpets whose thirst is satisfied by the chaotic pattern of water drops spouting fountain-like from an aperture at the apex of the sculpture.

Latterly, artists have been fascinated by the revelations of mental processes from modern imaging techniques, especially functional magnetic resonance imaging (fMRI). In one sense, art has always presented an experimental field for mental processes, but the nature of the processes themselves has now become accessible. Artists such as the English Andrew Carnie and the Scottish Annie Cattrell have seized upon the imagery emerging from neuroscience to create their own three-dimensional translations of the data generated by scientists with whom they



Light in space: in *Three Elements* by Nam June Paik, lasers endlessly rebound.

have collaborated^{8,9}. Cattrell’s series of *The Five Senses*, an ancient theme, translates the localization of sensory activity into translucent sculptures in which the jagged sites of transitory neuronal activity are captured forever within glassy cubes.

Those scientists who generate the data that so fascinate many artists are themselves frequently alert to the aesthetic dimension of their activity. This dimension can operate at the earliest levels of hypothesis formulation and mental modelling, and can intervene at any point in the processes of observation, experimentation and deduction that result in a publishable piece of science. Publication itself involves a series of conscious and unconscious choices of what the results should look like. It is virtually obligatory that any presentation of data in visual form should manifest a high-tech style. Many authors, particularly those aiming to communicate to an audience outside their immediate professional orbit, use artistry to stimulate engagement, impact and excitement. For many scientists, these dimensions are barely implicit, but for others the aesthetic motive is consciously present throughout. This is particularly apparent in astronomy and cosmology. Jean-Pierre Luminet’s recent reinstatement of a polygonal model of the cosmos — albeit one that exploits the strange spatial properties of a hypersphere — involves artistic instincts throughout, from conception to presentation.

So can art be science and science be art? Avoiding the pusillanimous answer that it depends on what kind of art and what kind of science, I should like to conclude by drawing attention not so much to differences in means that are sometimes clear, sometimes blurred, but to a divergence in intentional ends. A work of art always remains open for interpretation, drawing the spectator into the shape of the artist’s visualization, but without being able to exert fixed control over the feelings it induces. There is always room for the beholder’s share. Scientists may wish to engage the reader or spectator in a wonderful journey of imaginative visualization, but in the final analysis they wish to communicate an interpretation that embodies testable content in an unambiguous way. But behind these diverging streams of intention runs a turbulent river of shared intuitions about the order and disorder of things. ■

Martin Kemp is professor of the history of art at the University of Oxford, and co-director of Wallace Kemp Artakt.

1. Kemp, M. *Nature* **413**, 778 (2001).
2. Kemp, M. *Visualizations: The Nature Book of Art and Science* 1–4 (Oxford Univ. Press, Oxford, 2000).
3. Kemp, M. *Nature* **404**, 546 (2000).
4. Kemp, M. *Nature* **430**, 969 (2004).
5. Kemp, M. *Nature* **428**, 803 (2004).
6. Kemp, M. *Visualizations: The Nature Book of Art and Science* 144–145, 148–151, 154–155, 162–163 (Oxford Univ. Press, Oxford, 2000).
7. Kemp, M. *Nature* **419**, 882 (2002).
8. Kemp, M. *Nature* **416**, 265 (2002).
9. Kemp, M. *Nature* **424**, 181 (2003).

Experimental physics, experimental art

What happens when artists and particle physicists are brought together to exchange ideas?

Ken McMullen describes the creative fallout.

The concept of Signatures of the Invisible was sparked by a chance discussion I had with Maurice Jacob, high up in the Alps, on a snowy track in Briançon, France. It was December 1996, and I had got to know Maurice, then head of the Theory Division at CERN, Europe's particle-physics laboratory, through a social network entirely independent of science.

Rather than a movement, Signatures of the Invisible became a discourse between artists and particle physicists. It led to the creation of works of art that have been exhibited internationally from Beijing to New York, and have been seen by over a million people. In line with Signature's general philosophy, the works are not illustrative of science as such, but are responses — conveyed in various media and conceptual frameworks — to this discourse between artists and scientists. If Signatures of the Invisible had any manifesto at all, it was that it should not conform to any single artistic dogma.

Maurice is a man of wide cultural interests, and his generosity in taking time to describe to an outsider, an artist, what went on at CERN opened a door in my consciousness. At 1,500 metres above sea level, we conceived the idea of bringing together artists and particle physicists, and seeing what creativity would fall out.

"Artistic practice cannot derive from the artist's subjectivity alone, but demands engagement between the artist's intra-psychic processes and the external world."



Creativity ruffled: McMullen's work *Skin Without Skin* emerged from a discussion of crumple theory.

Within months of meeting, Maurice and I had arranged for this to happen at CERN in Geneva. A group of 12 leading European artists were offered free access to CERN's remarkable thinkers and experimenters for discussions about the strange new world of high-energy physics. It took another two years to organize funding, but following this, the 12 of us went to spend a week or so — some of us much longer — at the CERN laboratory. High levels of technical assistance helped us to produce works addressing what, as we

came to recognize through our discussions, are some of the most profound questions confronting people today.

Signatures of the Invisible was motivated by the belief that artistic practice, if it is to have meaning in the modern world, cannot derive from the artist's subjectivity alone, but demands engagement between the artist's intra-psychic (subjective or intuitive) processes and the external world in which we live. Until this point, little had been done by artists to address the insights offered by particle physics — quantum mechanics, relativity, matter and antimatter — into the underlying laws of nature.

Particle physics is of course one of the most demanding intellectual disciplines, and artists by their nature and training do not work in the same way. But artists have their own disciplines and emphases, their own ways of looking at things, which are equally valid and can offer corresponding insights into the nature of things. Take colour: for the artist it is a purely visual phenomenon, but for physicists it can relate to temperature, extending from the infrared to the ultraviolet and beyond. Take surface: for an artist, a surface is where you place things — paint on a canvas surface for example — while for the physicist, it is the outermost part of a body, and may be associated with levels of energy. Take units of time: for the film-maker it is 24 frames per second; for a physicist it may be a matter of femtoseconds.

Signatures of the Invisible was not, then, a meeting of opposites, but a correspondence between equal partners from different traditions. It was not a straightforward illustrative 'sci-art' project, but a free flow of ideas between groups of people, each concerned in their own way with the mystery and the reality of what makes us who we are; each trying to understand the Universe in which we find ourselves a part.

During the three years of engagement with CERN, I found myself in an unknown world, which while offering a unique opportunity to explore new territory, also challenged many previously held certainties. The journey was exciting and at times terrifying as far as my own artistic practice was concerned. There were many periods when I was completely lost. There were also moments of exhilaration when some works found their form and occasionally their resolution.

At one point I found myself in free-wheeling discussions with John March Russell, a CERN theorist, about the physics involved in crumpling a piece of paper: I discovered there is actually a 'crumple theory'. We began to talk of the surface as a concept, the kind of energy that is released in the act of crumpling, and the strange fact that a crumple can never be repeated exactly — even if the paper is the same construction right down to the atomic level, and the same energy and force are applied. So the crumples are probably a concrete signature of time. Fascinated, I continued these discussions at highly technical levels with CERN



A mark of time: the crumple on the right became the sketch for *Skin Without Skin*.

engineers who helped me to experiment for months with the theory: crumpling many different pieces of paper, and eventually crumpling giant sheets of titanium. We used laser scanning, laser cutting, laser welding — high-tech tools that I would otherwise never have come to use in my work.

The Signatures philosophy carried the risk of complete failure, of course, and some ideas did fail. But the exhibition that Signatures

spawned, of over 50 diverse artworks, showed the overall success of this encounter. Several of us are now engaged in vigorous exploration of dark matter and dark energy, and the consequences of these discoveries, at the Stanford Linear Accelerator Center (SLAC) in California.

Ken McMullen is an artist and film-maker, and research professor at London's University of the Arts.

TAPE TRANSCRIPT FROM A CERN ENCOUNTER

John March Russell If I think about the interaction of two particles, I think of two lines meeting... There's this strange thing called 'a space-filling curve'... You think of space being bigger than a line or a surface being bigger than a line but actually... in a sense they have the same number of points.

Ken McMullen I like that... I like that a lot.

JMR If I draw a plane and count the number of points on that plane, there are actually the same number of points as if I'd drawn a line... This seems counter intuitive but both are actually infinite. A space-filling curve does crazy things... it's so bendy that it fills up the surface or volume completely.

KM The line would have to cross itself though.

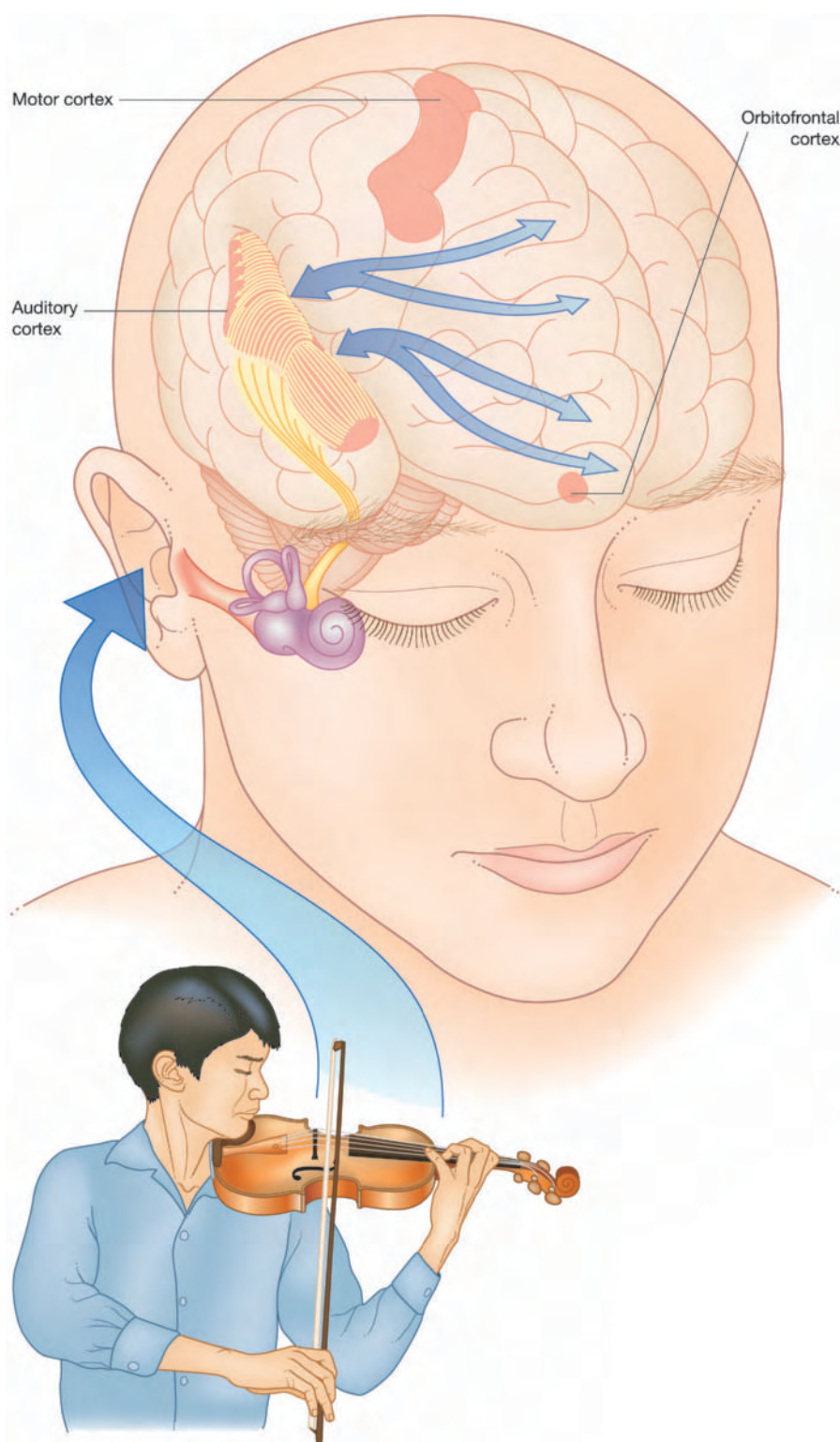
JMR No, you can do this mathematically... It twists round in a very complicated way... There are people here who could demonstrate this.

KM If we do these surfaces with lines on them we can trace the lines in such a way that in the end — this is a technical question — all that would exist...

JMR ... is the lines.

KM It would be a strange set of lines because they would have been infiltrated with the crumpled surface.

JMR It should be possible... who knows what it would look like.



Music, the food of neuroscience?

Playing, listening to and creating music involves practically every cognitive function. Robert Zatorre explains how music can teach us about speech, brain plasticity and even the origins of emotion.

We tend to consider art and culture from a humanistic or historical perspective rather than a biological one. Yet these products of human cognition must have their origin in the function and structure of the human nervous system. As such, they should be able to yield valuable scientific insights. This line of reasoning is nowhere more evident than in the contemporary interest in the neuroscience of music.

Music provides a tool to study numerous aspects of neuroscience, from motor-skill learning to emotion. Indeed, from a psychologist's point of view, listening to and producing music involves a tantalizing mix of practically every human cognitive function. Even a seemingly simple activity, such as humming a familiar tune, necessitates complex auditory pattern-processing mechanisms, attention, memory storage and retrieval, motor programming, sensory-motor integration, and so forth (Fig. 1).

Likewise, the musician does not consider music to be monolithic, but recognizes within it multiple features including melodies, chords, themes, riffs, rhythms and tempos. This complexity — both psychological and musicological — makes music a challenging topic for a scientific research programme. Increasing numbers of investigators are convinced that music can yield valuable information about how the brain

Figure 1 The processing of sound waves from a musical instrument. After being transduced into neural impulses by the inner ear, information travels through several waystations in the brainstem and midbrain to reach the auditory cortex. The auditory cortex contains distinct subregions that are important for decoding and representing the various aspects of the complex sound. In turn, information from the auditory cortex interacts with many other brain areas, especially the frontal lobe, for memory formation and interpretation. The orbitofrontal region is one of many involved in emotional evaluation. The motor cortex is involved in sensory-motor feedback circuits, and in controlling the movements needed to produce music using an instrument.

works: they believe that the study of the brain and the study of music can be mutually revealing.

How does one go about studying this intricate thing called music? Few scientists would accept that such a complex function could be studied, let alone understood, without first identifying and describing its various components. But this raises the thorny problem of deciding which components of music are pertinent, and how these components are shared or distributed among different cognitive functions. Some cognitive functions, such as figuring out pitch interval ratios, may be unique to music, whereas others, such as memory, may be general systems that are used in many different domains.

The oldest scientific technique for understanding brain functions is to study the consequences of brain lesions. We have long known that severe damage to the auditory cortex — where information coming from the ear is first analysed and interpreted — disturbs the ability to make sense of sounds in general. But occasionally, lesions of certain auditory cortical regions result in an unusual phenomenon: a highly selective problem with perceiving and interpreting music, termed ‘amusia’¹.

People with this type of damage have no problem speaking or understanding speech, or making sense of everyday

sounds. But they cannot notice wrong notes inserted into tunes, or recognize even the most familiar melody. Even more surprising is that a minority of otherwise normal individuals appear to be born with the same inability to recognize tunes. In some cases, the deficit seems to run in families, suggesting a genetic component².

This extraordinarily selective problem in processing music, whether acquired or inborn, could result from very selective damage or dysfunction in an area of the auditory cortex where fine-grained pitch differences and sound frequency ratios (musical intervals) are processed³. Such a specific deficit at one of the earliest steps of music processing could propagate through the perceptual system, resulting in a global disability. The ability to compute pitch relations is critical to music processing, and if the brain is unable to represent pitch, the entire music perception mechanism could easily be destabilized.

The study of people with amusia has shown us that music depends on certain types of neural process. Such people provide living examples of what results when these neural processes are disrupted. And they have shown us that music can indeed lend itself to scientific study.

Music and speech

Scientists would also like to understand why we have evolved a sense for music in the first place, and, in particular, whether musical ability is somehow an extension of speech: many have argued for this on the reasonable grounds that music and speech share several formal similarities. So researchers have tried, using various techniques, to determine the extent to which the processing of music and that of speech share neural resources. The results so far are somewhat conflicting, but also intriguing.

One of the striking things about the neurobiological processing of speech is that it mostly takes place in the left half of the brain. It has therefore been natural to ask whether

this asymmetry is mirrored in a right-hemisphere predominance for music. There are also many case reports of individuals who have lost their speech functions after extensive damage to speech regions in the left cerebral hemisphere, yet continue to show intact high-level

musical function (for example, the Russian composer Vissarion Shebalin⁴).

These data suggest that music and speech processing do not use completely overlapping neural substrates. But neuroimaging studies indicate that some functions, such as syntax, may require common neural resources for both speech and music⁵. In other words, the ability to organize a set of words into a meaningful sentence and the ability to organize a set of notes into a well-structured melody might engage brain mechanisms in a similar way.

But the data from which we have drawn these conclusions have limitations. On the one hand, many of the case reports were studied in a descriptive, anecdotal manner. On the other hand, neuroimaging can be notoriously difficult to interpret: similar patterns of brain activity do not necessarily mean that similar neural substrates are involved, because many complexities of

neural patterning are beyond our present technology’s ability to measure.

The key to resolving these questions comes from a more systematic understanding of the different cognitive components involved, and the specific neural circuits associated with them. Fine-grained pitch processing — a highly critical component of music perception — has proven particularly valuable in dissecting the differences between how the brain handles speech and music.

Recent evidence from functional brain activation, magnetic recording and lesion studies, suggests that a particular region of the auditory cortex in the right hemisphere is much more specialized for representing detailed pitch information than its counterpart on the left side of the brain. Tones that are close together in pitch seem to be better resolved by neurons on the right.

Why should this functional segregation have emerged? It could be related to the requirement to sample sound information from the environment in different ways, according to need: either quickly and roughly, or if time allows, accurately⁶. If the sound energy is changing very rapidly, for example, a quick snapshot may be needed. The perceptual system needs to track these changes online, and hence must sacrifice detail to achieve speed. Such may be the case for speech recognition where detailed temporal information is essential to recover the sounds produced by the rapidly moving articulatory muscles of the lips and tongue. Conversely, some aspects of sounds that are important for perceiving music evolve much more slowly, so the nervous system can take a more detailed look at the structure of the sound. This takes more time, of course, but yields a finer-grained internal representation. Naturally occurring periodic sounds (many vibrating objects, voices or animal calls) contain pitch information that is important to process. Pitch is also a good cue to distinguish one sound from another in a noisy environment. So the postulated pitch processing mechanisms need not have evolved for music *per se*, but could be part of a general system for using natural sounds from the environment.

Thus, the different specializations of the auditory cortex on the two sides of the brain can be seen as different parameter settings on what are essentially two parallel systems. This approach shows us that it is perhaps less

“Art and culture must have their origins in the function and structure of the human nervous system, and should be able to yield valuable scientific insights.”

Perfect pitch: children can acquire absolute pitch only if they receive musical training before the age of 12 to 15.

interesting to ask, “on which side of the brain is music processing located?” than to set about systematically studying the various subcomponents that contribute to various aspects of musical function.

Music and development

Another reason music has caught the attention of scientists trying to understand the brain is that the ability to perceive music seems to be present from very early in development. Of course, we learn the specifics of our musical culture from the environment. But the human infant seems to come into the world with a brain already well prepared to figure out its musical world.

Any mother can attest to the way an infant will respond to the pitch and rhythm of her voice. But babies are surprisingly sophisticated mini-musicians: they are able to distinguish different scales and chords, and show preferences for consonant over dissonant combinations, for example⁷. They can recognize tunes played to them over periods of days or weeks, and are capable of remarkable feats of statistical learning, being sensitive to regularities in sounds⁸. In other words, babies’ nervous systems seem to be equipped with a capacity to sort out the different musical sounds reaching their ears in order to construct a grammar, or system of rules.

It could be argued that this is part of a general capacity to make sense of the world — to be able to predict what is coming up next. In some sense this is certainly true. But the notable thing is that this ability endows infants with the capacity expressed later in life to respond to and enjoy music. All this evidence supports the general idea that the ability to perceive and process music is not some recent add-on to our cognition, but that it has been around long enough to be expressed from the earliest stages of our neural development.

Music involves not only listening, but also playing and creating, where individual differences are much more evident. Although nearly everyone seems to have sophisticated neural systems that allow them to perceive music, and to reproduce musical patterns by singing, not everyone is able to play the piano like Vladimir Horowitz.

This leads to two very interesting scientific questions, which are the subject of active research. How can we explain individual differences in ‘native’ ability? And what effects does training have on brain function and structure? Little progress has been made on the first question, except in the very specific domain of ‘absolute pitch’, where interactions between genetic and environmental factors are beginning to be unraveled⁹. It is now clear that absolute pitch cannot develop without some musical training, but critically, the exposure must happen during childhood: past the age of 12 to 15, it is essentially impossible to learn it. From this one can conclude that the brain must be particularly sensitive during a certain time in development. But not all children given music lessons develop this skill, so other factors must also be at play. New evidence suggests that genetics has a role¹⁰. This is a field to watch in the near future.

“Findings of brain plasticity have very general implications for our understanding of the interplay between the environment and the brain.”

In contrast, a number of very clear findings are now emerging that help us to understand how the brain is sculpted by musical experience. Most of this work shows that training in music enhances the activity of certain neural systems. For example, areas of the motor cortex corresponding specifically to the fingers of the left hand show an enhanced electrical response among violin players¹¹. These changes are directly related to the age at which training is begun: those who began studying music in early childhood show the most extensive modification to brain response, whereas those who waited until after puberty show much less. Similar effects have been described for the auditory cortex’s response to sounds produced by specific instruments¹¹.

Moreover, anatomical changes accompany these enhancements in responsivity. Several studies have reported greater tissue density, or enlargement of motor- and auditory-related structures among musicians, indicating that

years of training actually change the underlying structure of the nervous system¹². These findings should not be taken as evidence that music makes a person’s brain bigger and therefore better. The changes are very specific, and it could be that they come at the expense of other functions. But such findings of brain plasticity have very general implications for our understanding of the interplay between the environment and the brain, particularly in the context of development, as the age at which training takes place is so critical.

Music and emotion

One of the questions that most frequently comes up in discussions of music, and yet has received relatively little attention in the neuroscience community, concerns emotion. Indeed, non-scientists are often puzzled that this aspect has been relatively neglected in favour of more esoteric concerns, given that, for most people, music exists solely to express or communicate emotion. There are some sophisticated treatises in the musicological tradition on this question (for example, the classic volume by Leonard Meyer¹³), but only recently has the topic begun to attract serious attention from neuroscientists¹⁴.

One thing we do know is that music can elicit not only psychological mood changes, but also physiological changes in heart rate, respiration and so forth, that mirror the changes in mood. Indeed, music’s anxiolytic effect is known not only to the specialist, but to anyone who listens to a favourite piece of music to relax after a trying day.

What brain responses can explain these effects? At the moment we simply don’t know. But plausible hypotheses are guiding research. One notion is that music results in physical entrainment of motor and physiological functions: music drives the body. So, loud, rhythmic, fast music tends to make you feel lively — or even want to dance — whereas slow, soft music leads to calmness, and even sadness. A possible explanation is that these effects could be mediated through sensory–motor feedback circuits, which have been much discussed in neurophysiology; that is, through the so-called mirror-neuron system¹⁵. Although there is no direct evidence for this idea, it is plausible in that this system is thought to mediate imitative behaviour by linking perception directly to action. A

similar mechanism might explain some of the effects of music on physical movement, and so mood induction.

But music's emotional undercurrents run deeper than such an analysis might suggest. Studying the very complex and idiosyncratic responses to music is challenging because it depends on so many difficult-to-control factors, not least individual preferences. What is 'music' to one person's ears is often offensive to another's (consider teenagers and their parents as a typical example). So cultural and social factors clearly have important roles in modulating our emotional response to music. Yet there are still likely to be common neural pathways that mediate responses, such as pleasure, to music.

One intriguing and very specific emotional response is the 'chills down the spine' effect. Anyone who has experienced this knows exactly what I refer to: for the minority who haven't, it won't do much good to try to explain it. But we are beginning to understand some of the neural mechanisms that underlie these kinds of response. When listeners experience the chills, neuroimaging shows that the brain areas recruited include regions thought to be involved in

mechanisms of reward and motivation. Examples are the basal forebrain and certain brainstem nuclei, along with cortical areas involved in emotional evaluation, such as the orbitofrontal and insular regions¹⁶. These circuits are similar to those involved in mediating responses to biologically rewarding stimuli, such as food or sexual stimuli.

"Maybe music, and all art in a way, manages to transcend mere perception precisely because it contacts our more primordial neurobiology."

But why should music, an abstract pattern of sound, have any commonality at all with such survival-related systems? It is a stretch to suggest that music is essential for life or reproduction. However, perhaps this research is beginning to illuminate the complex relation between cognitive-perceptual systems that analyse and represent the outside world, and evolutionarily ancient neural systems involved in assessing the value of a stimulus relative to survival and deciding what action to take. Maybe music, and all art in a way, manages to transcend mere perception precisely because it contacts our more primordial neurobiology.

To caricature the idea, we can think of the neocortex as being able to analyse relations and notice patterns, but then this processed information interacts with the emotion/evaluation system, which in turn leads to pleasure

(or sadness, fear, excitement and so forth). The vagueness of these concepts indicates how far we are from having anything like a model of the processes going on — although an optimist might point out that even being able to talk about it, albeit in unclear terms, shows how far we have come. ■

Robert Zatorre is a cognitive neuroscientist and James McGill professor of neuroscience at the *Montreal Neurological Institute*.

1. Ayotte, J., Peretz, I., Rousseau, L., Bard, C. & Bojanowski, M. *Brain* **123**, 1926–1938 (2000).
2. Drayna, D., Manichaikul, A., de Lange, M., Snieder, H. & Spector, T. *Science* **291**, 1969–1972 (2001).
3. Peretz, I. *et al. Neuron* **33**, 185–191 (2002).
4. Luria, A., Tsvetkova, L. & Futer, D. *J. Neurol. Sci.* **2**, 288–292 (1965).
5. Patel, A. *Nature Neurosci.* **6**, 674–681 (2003).
6. Zatorre, R. J., Belin, P. & Penhune, V. B. *Trends Cogn. Sci.* **6**, 37–46 (2002).
7. Trehub, S. E. *Nature Neurosci.* **6**, 669–673 (2003).
8. Saffran, J. R. in *The Cognitive Neuroscience of Music* (eds Peretz, I. & Zatorre, R. J.) 32–41 (Oxford Univ. Press, New York, 2003).
9. Zatorre, R. J. *Nature Neurosci.* **6**, 692–695 (2003).
10. Baharloo, S., Service, S., Risch, N., Gitschier, J. & Freimer, N. *Am. J. Hum. Genet.* **67**, 755–758 (2000).
11. Pantev, C., Engelien, A., Candia, V. & Elbert, T. in *The Cognitive Neuroscience of Music* (eds Peretz, I. & Zatorre, R. J.) 382–395 (Oxford Univ. Press, New York, 2003).
12. Schlaug, G. in *The Cognitive Neuroscience of Music* (eds Peretz, I. & Zatorre, R. J.) 366–381 (Oxford Univ. Press, New York, 2003).
13. Meyer, L. B. *Emotion and Meaning in Music* (Univ. Chicago Press, Chicago, 1956).
14. Juslin, P. N. & Sloboda, J. A. *Music and Emotion: Theory and Research* (Oxford Univ. Press, Oxford, 2001).
15. Rizzolatti, G. & Arbib, M. *Trends Neurosci.* **21**, 188–194 (1998).
16. Blood, A. J. & Zatorre, R. J. *Proc. Natl Acad. Sci. USA* **98**, 11818–11823 (2001).

The evolution of sensibility

Understanding the science behind aesthetic perception could guide and restrain the 'shock of the new' approach to music. Composer Roger Reynolds explains how.

Let us assume, at the outset, that artists have what English refers to as a 'gift' — a knack for making certain connections between observation, experience and action. At first, I think, one's gift is shaped by example: what one experiences, whether by inadvertence or intention. Later, the opinions of particular individuals may acquire weight as well. One entertains things that were not at first appealing. And one comes to realize that value resides not only in immediately palpable sensation, but in the shape of an experience, what is referred to, how a structure articulated in the now relates to past habit and memory.

There is a period when one is uncertain about the nature of aesthetic value: one is suspended between that which directly

appeals and that for which merit is asserted on more abstract grounds. But over time, one learns how to balance the authority of the intellect with that of immediate intuition. One begins to clear paths, the boundaries and directions of which depend on the aesthetic sensibility that one is discovering. What was 'negotiable' and open to external influence and internal debate becomes a matter of fact. One knows what one can accept as art. There is no need for the time-consuming process of weighing one evaluation against another. Consideration is not precluded: it is simply that the nature of one's engagement with aesthetic acts alters over time. And a strong level of certainty tends to develop regarding what is and what is not acceptable.

Aspirations

The artist's central compulsion is to invent, to hear what has not been heard, to see what has not been seen — yet to do these things within a context that is not marred by the inappropriate. This quest is not simply driven by a need to be different. It is rather the result of a thirst for the gratifications of that which is novel: a new sound, a new connection between known components, a new course set out upon. Delight comes from the thing itself — its proportions, its tactility, the way it refers to what one has already known. But art also covets the sense with which its particular moments put one in touch; a sense of the fresh, and even, in a limited way, the unprecedented.

But what of the needs of those who receive aesthetic experience rather than create it? Novelty and sensual appeal have attractions here too; but comprehension, ratification and the reinforcement or denial of expectation are important as well. A listener may occasionally be open to experiencing works that are perplexing (even seemingly indifferent to comprehension). But in such instances there is an unspoken assumption that they will ultimately be able to grasp the whole.

So, in brief, the composer's aspiration is to acquire the unknown, while the listener desires the ratification of that which is already known (or, at the least, knowable). In recent decades, the 'space' of that which might be explored — the possible musical vocabularies, syntaxes and strategies that may attract composers — has expanded at a seemingly undamped and possibly undampable rate. Meanwhile, opportunity



Roger Reynolds teamed up with perceptual psychologists to test the validity of his composing strategies.

T. MARTINOT

(in the shape of commissions and performances) has increasingly been dictated by the economic pressure to traffic in the easily recognized and immediately appealing. This is an unfortunate union. Science, which has driven the expansion of opportunity by developing and proliferating new technologies, should perhaps have a role in ameliorating the impact of these changes.

Considerations

For artists living at earlier stages in the evolution of Western music — especially in relatively slow-evolving ‘common-practice’ periods — widely accepted (and unchallenged) consistencies of behaviour held sway. This is not the contemporary situation. A significant proportion of the music written since the Second World War exemplifies unprecedented premises. The most notorious of these include the desirability of using all sounds, not only those that have been thought ‘musical’; the belief that ‘intention’ is best disavowed, and replaced by chance procedures that distribute events over time in a fashion that is distanced from purpose; and the belief that, because differences between sounds can be detected, these differences can be relied upon to carry musical information.

Although I was trained, and I matured, during one of the most turbulent periods for the arts in recent memory (the 1960s), my experimental nature was buttressed by formal training in a path-breaking engineering physics programme at the University of Michigan. This training caused me to respect planning, and question, from a variety of perspectives, ideas or experiences. I embraced the notion (no mentor argued for it) that musical experience would benefit from foresight, from explicit planning. My rather late coming to music as a serious student (at the age of 14) meant I was more likely to consider the necessity of something than to acquire it as unexamined habit.

I was immediately attracted to the work of innovative composers, not only because of the sonorous persuasiveness of the work itself, but also because such composers explained their purposes in arresting ways. I think, in this regard, of Charles Ives, Edgard Varèse, John Cage and Iannis Xenakis. It is not incidental that several of these were strongly attracted to science, as I was. Some, such as Varèse, used the fruit of scientific explorations as inspiration

in a poetic or metaphorical sense. Others, particularly Xenakis, who was educated at the Athens Polytechnic, used explicit and detailed computation; for example, to emulate stochastic process in sound through complex orchestral textures. But all these explorers shared the habit of asking unsettling, sometimes deliberately iconoclastic questions. Their resistance to the assumption that the way things had been done was necessarily the way things should continue to be done influenced my course considerably.

Western musical notation is a highly evolved and efficient system, but it is suited to eliciting well-established categories of sonic behaviour. In recent decades, composers have sought more flexible ways to describe musical activity. This along with the emergence of computers and an allied interest of composers and computer programmers in psychoacoustics, has resulted in more general and more abstract graphic representations becoming attractive and useful.

Because of my engineering training, I was not put off by this trend to capture ideas and the behaviour of new sounds using graphic strategies that are far removed from the conventional stave-based representations of classical music. I gradually forged an amalgam of practices that was variously indebted: to what spontaneously moved me; to what interested me intellectually; and to what those whom I admired espoused. With so many influences (which were by no means mine alone), the optimization over time of one’s approach to compositional technique and practice is unlikely to cohere naturally. There are simply too many component threads in the weave. It is difficult to imagine how someone living in the slow-moving ‘common-practice’ periods of Western music could have had anything but coherent premises. Now, it is far more likely that self-contradictory, or at least inadequately examined, premises will arise within an artist’s ways.

As a member of the international arts community, who travels frequently and constantly encounters new work, I find much of what I hear unsatisfying. As previously noted, once one has reached a certain stage of development, one is no longer concerned with whether something is an ‘acceptable aesthetic experience’. One tends to behave as

though one’s belief or practice is simply a matter of fact. But, of course, this force of belief does not reliably extend beyond the limits of one’s own internal world.

Because of my science training, it was clear to me that unproved assertions needed to be tested in some way. But the tendency in the contemporary music world of the 1960s was to equate the value of the ‘spontaneous’ with that of considered action; to assume that the vigorous assertion of something (especially if it was accompanied by abundant heat and jargon) was sufficient to establish it as viable practice. There were some

notable combatants: Xenakis resisted Serialism (the twelve-tone system of music introduced by Arnold Schönberg in the early twentieth century, and elaborated by Pierre Boulez and others in the 1950s); and the Serialists resisted the incursions of chance or indeterminate procedures pioneered by John Cage. But eventually, examinations of the assumptions underlying a given practice was resisted and was even considered uncollegial. Live and let live.

Several years ago, a number of experiences in my life as a composer, lecturer and mentor forced me to realize that I would have to address my deepening concerns. I offered graduate seminars in which assumptions were critically examined, wrote on my own practice, and ultimately became involved in a multi-year collaborative project with perceptual psychologists. My perspective was more limited and introspective than that of the psychologists. I hoped that together we could establish, with some degree of objective confidence, whether the strategies I had developed during four decades of composing were perceptually plausible. My colleagues, in turn, wanted to evolve strategies for examining and understanding the experience of musical form.

Several years ago, a number of experiences in my life as a composer, lecturer and mentor forced me to realize that I would have to address my deepening concerns. I offered graduate seminars in which assumptions were critically examined, wrote on my own practice, and ultimately became involved in a multi-year collaborative project with perceptual psychologists. My perspective was more limited and introspective than that of the psychologists. I hoped that together we could establish, with some degree of objective confidence, whether the strategies I had developed during four decades of composing were perceptually plausible. My colleagues, in turn, wanted to evolve strategies for examining and understanding the experience of musical form.

The Angel of Death Project

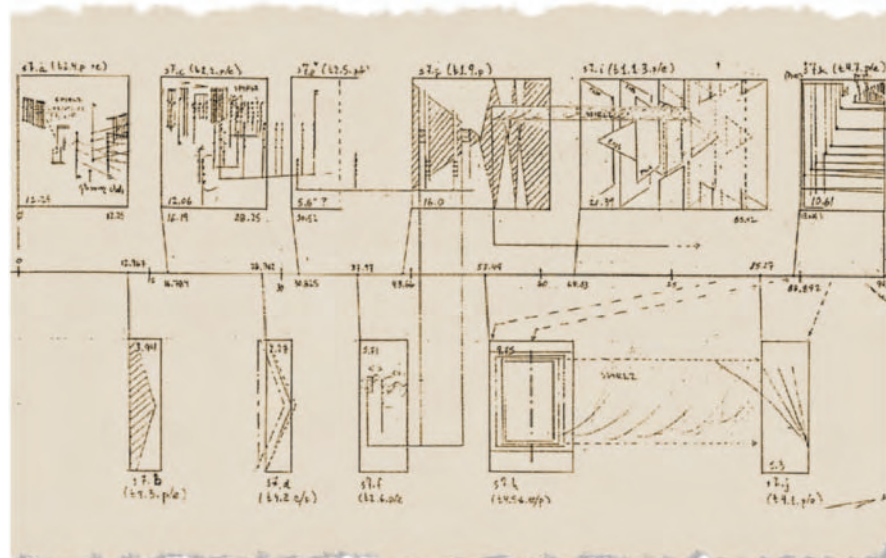
Stephen McAdams, the long-time head of the Music Perception and Cognition Team at the French music research facility, Ircam, contacted me in 1993. He proposed that I collaborate with his team to explore how listeners respond to music during concerts. Most studies of the meaning of large-scale form in music — what a listener perceives

“The composer’s aspiration is to acquire the unknown, while the listener desires the ratification of that which is already known.”

and experiences during the complete span of a performed piece of music — have been theoretical and speculative. Perceptual studies on musical experience have been largely conducted using brief tunes or harmonic progressions created, in turn, within the limits of traditional tonal assumptions. McAdams wanted to enlarge this envelope dramatically by examining in-concert listening experience: he wanted to probe into whether (and if so, how) a composer's pre-suppositions about his work might be confirmed or discredited by formal experiments with musically trained and also naive listeners. These studies were also designed to address music that was not founded upon the traditional conventions of musical keys, their characteristic harmonies, the classical forms.

I composed a piece, *The Angel of Death*, that was explicitly suited to becoming an experimental object. It is in two parts that have strong underlying similarities in structure and subject but that differ markedly in their natures. The 'sectional' part is constructed as a form within which there are clearly distinct sections, each containing its own correspondingly characteristic norms in the deployment of musical elements within the section's boundaries. The 'domain' part is more organic. In this part, the deployment of the same musical elements is continuously evolving, which tends to de-emphasize boundaries. The work also features two complementary media: a solo pianist and a chamber orchestra. Thus, a particular musical function — the presentation, say, of a theme — can be assigned to either the soloist or the ensemble. The roles of the soloist and orchestra reverse from one part to the other, and either the domain or the sectional part can begin a performance. The modularity and variability of the components of *Angel* could allow perceptual psychologists to test assumptions in a variety of ways that would not be possible using a more traditionally conceived composition.

Brief descriptions of two experiments will illustrate their approaches. I used five themes in fashioning the primary parts of my composition. Each theme is very brief and was conceived as a complete statement



Musical invention: a planning sketch for the final computer section in *The Angel of Death* showing the nature and interactions of its components.

in miniature, positing and then traversing a musical landscape of its own. The themes each have subsections, which were represented by 34 buttons on a computer screen — one for each subsection. Clicking on a button allowed the listener to hear the corresponding fragment. The task was to drag the buttons into groups of 'like material'. The experiment probed listener comprehension of the similarities and distinctions

"I hoped that we could establish whether the strategies I had developed during four decades of composing were perceptually plausible."

between my thematic elements, and revealed the ways in which these perceptions confirmed or challenged what I had intended to do in the composition.

A second, more innovative study evaluated how two criteria fluctuated for listeners hearing the work in a concert setting. These criteria were 'familiarity' and 'emotional force'. Familiarity was the degree to which materials at a given moment were felt to be familiar or novel compared with anything heard from the beginning of the piece. Emotional force indicated the level of arousal felt (whether positive or negative). Responses were recorded by listeners manipulating sliders on their individual boxes as the performance proceeded. Average scores were then mapped on to a graphic representation of the formal structure of

the work as it had been conceived and composed. The purpose here was not only to examine the differences and commonalities between the individual-response curves, but to see whether listener perceptions confirmed the existence of formal landmarks predicted by the musical form as conceived.

The results of these and other studies have been published^{1,2}. In general, they confirm my hope that the strategies I evolved during 40 years of experience are perceptually plausible. As indicated, I based the musical discourse in *Angel*'s two contrasting parts on a small set of brief themes. Were these themes so diverse in their natures that listeners were unable to hear familial similarities, or — perhaps worse — were the materials so undifferentiated that listeners were unable to detect differences between them, my dramaturgical prospects would be hobbled. (Substitute the word 'character' for 'theme' and imagine the effect of inappropriate differentiation on the story line in a novel.) What the experiments showed was that listeners — whether musically sophisticated or not — heard the materials I had written in ways that I judge almost ideal. Their characterizations make complete sense, if one keeps in mind the fact that certain kinds of ambiguity are essential to the purposes of art.

Loud and clear: few perceptual studies on musical experience have looked at in-concert listening.

For in-concert listening, the statistical fluctuation of listener response (whether in terms of the novelty of materials or the emotional force aroused) was graphically correlated with the composed score and the audio recording: the fluctuations in response were compared with what my score identifies as significant 'landmarks' in the musical design. When listeners respond significantly as a group (their responses wax or wane in magnitude), these graphic features are always associated with aspects of the musical form that I have described as important.

On a less favourable note, the data also have clear implications about the difficulty of the cognitive tasks provided by music whose premises have not been well inculcated into a listener's mind. One experiment, for example, explored the effects of presenting identical sets of musical relationships — specific patterns of pitch and rhythm — and weighed the effects of presenting them through different media (orchestra instead of solo piano, for example). When the idiom is unfamiliar, a repeat presentation of the 'same' musical fact is poorly recognized if the auditory medium which carries it changes. The tasks presented

by non-traditional music seem considerably more demanding than many contemporary musicians would like to believe.

Integrity

In spite of abundant opportunity — in terms of stylistic variety, technological enhancement, multimedial combinations, and so on — the contemporary artist nevertheless has a relatively small domain of choice within which to navigate. One's integrity is enforced by one's aesthetic sensibility. Certain things are acceptable or desirable; others are not. One is obliged to be what one is — or from a slightly more circumspect perspective, to be what one has become. The degree to which

"The aims of the creator and his or her audience have got markedly out of phase with one another."

the artist's worlds of personal imperatives conform to the facts of human perception and cognition is crucial to the fate of our art. As I suggested at the outset, the aims of the creator and his or her audience have got markedly out of phase with one another. This need not be attributed to perversity on the part of the artist or to recalcitrance on the part of the listener. It is rather a reflection of their differing needs in relation to aesthetic experience, and of the unprecedented pace at which our experience is reformulated by the rush of events.

The forces that propel the evolution of imagination and its application to aesthetic purpose show no sign of losing strength. So we cannot reasonably anticipate a period of calm and measured retrenchment any time soon. Because art is the product of individuals who share the same experiential world as their audience, it is highly improbable that what art tells us is irrelevant. The artist's charge is to process experience and to act in response to the life within which he or she is embedded. The rub lies in whether the ways that musical material is conceived and manipulated conform sufficiently to the processes by which our physical and psychological faculties inform us about the world.

I see no implication in this argument that art itself or the ways in which it changes as time and circumstance alter will or should be abridged. The opportunities for artists (and thus, for lovers of art) are uniquely extensible in our era. But it seems that there is — by virtue of the rapidity and drama of the science-driven changes of our time — an urgent need to ensure that the speculative fancy of the artist is informed by extensive and accurate information regarding the perceptual and cognitive capacities of the human being. Collaborative projects such as The Angel of Death Project should be undertaken by other artists, and by artists other than musicians. It will not be straightforward, but there is a need for compensatory processes. The effects of rapid change might be ameliorated by information that helps, for example, a composer, to know not what should be done, but rather what might be done. One hopes that such collaborations would do more than simply demonstrate the stylistic predilections of their authors. We should not seek the ratification of established perspectives, but a clear picture of potential prospects. Art needs to be free of inappropriately constraining presuppositions. On the other hand, its practitioners must remember that art addresses human beings, with all the scope and also the limitations that this implies. ■

Composer Roger Reynolds is professor of music at the Department of Music, University of California, San Diego, La Jolla, California 92037, USA.

1. Levitin, D. J. & Cuddy, L. L. (eds) *Music Percept.* 22 167–364 (2004).
2. McAdams, S. & Battier, M. (eds) *Perception and Creation of a Contemporary Musical Work: The Angel of Death Project* by Roger Reynolds, IRCAM-Centre Pompidou, Paris (in the press).

Poetry and science: greatness in little

Peter Forbes explores how poetry and science can mutually inspire.

I wrote this essay concurrently with doing research for a book on biomimetics. Two poems lived with me constantly during this process and provided epigraphs to most of the chapters. They exemplify one way in which poetry can engage with science.

The poems are *Greatness in Little* by the seventeenth-century English poet Richard Leigh, and *De Rerum Natura* by the first century BC Roman poet Titus Lucretius. What Leigh and Lucretius had in common was a curiosity that burst the bounds of existing knowledge: they wanted to know what lay beyond human senses and received wisdom. Their works were prescient because they happened to light on areas of science that have been elucidated since. But even if we withhold praise for their getting it right, we must be moved by their missionary zeal. The poetry lies not especially in the poets' technique, although Lucretius is immensely readable in translation: it lies more in the subject matter and the blazing ardour brought to bear on it.

Leigh's poem is less well-known than Lucretius' and, at 60 lines, is a lot shorter than Lucretius' 8,000 plus. For me, the great attraction of *Greatness in Little* is that its subject seems to be that in which I am most interested: biomimetics.

The title foreshadows Richard Feynman's famous 1959 rallying call for physicists to enter the new field of nanotechnology: "Plenty of room at the bottom." At the time Leigh was writing, the microscope was beginning to reveal the microworld of nature — a nature that "... shuts up her minuter bodies all / In curious frames, imperceptibly small".

To a biomimetician, Leigh's lines could be seen as a manifesto for the science:

Poet-cum-scientist: in his writing Primo Levi conveyed his early excitement about the natural world.

*What Skill is in the frame of Insects shown?
How fine the Threds, in their small Textures spun?
How close those Instruments and Engines knit,
Which Motion, and their slender Sense transmit?
Like living Watches, each of these conceals
A thousand Springs of Life, and moving wheels.*

He even seems to describe the nanostructures themselves:

*Those rugged little bodies whose parts rise
And fall in various inequalities,
Hills in the risings of their surface show,
As valleys in their hollow pits below.*

Leigh, you feel, would have been very happy to have been alive at this hour. His poem concludes: "These Atom-Worlds found out, I would despise / Colombus, and his vast Discoveries".

The pleasure of reading a poem such as Leigh's is one of shared appreciation of nature's mysteries. We are not going to learn any science from a poem but it is good to sometimes talk about science in elevated language. Otherwise, to paraphrase an almost forgotten poet, John Wain, we "grow yellow on a diet of prose". Ancient writing brings the added pleasure of seeing something eerily half-alive before its time.

Lucretius affords the same enjoyment as Leigh but in greater abundance. His poem *De Rerum Natura* is an epic that sets out, with

great passion and fine reasoning, the atomic theory of Democritus and Leucippus. Some of his lines always seem relevant to whatever you happen to be doing at the time. Working on structural colour in nature? Here is Lucretius:

*...first-beginnings have no colour,
But they do differ in shape, and from this cause
Arise effects of colour variation...
Hues change as light fall comes direct or slanting...
A peacock's tail, in the full blaze of light,
Changes in colour as he moves and turns.*

His arguments are ingenious and the poem contains, among other examples of prescience, a foreshadowing of Galileo's and Newton's work on motion, a convincing argument for indeterminacy and — in one of his most evocative passages — a prediction of brownian motion.

*...If you look sometimes,
You see the motes all dancing, as the sun
Streams through the shutters into a dark room.
Look! — there they go, like armies in maneuver
Whose little squadrons charge, retreat,
join, part,
From this you can deduce that on a scale
Oh, infinitely smaller, beyond your sight,
Similar turbulence whirls.*

Lucretius reminds us that, although in the Anglo-Saxon world the traffic between

literature and science has been fitful, in Italy there has been a strong tradition of exchange. Begun by Lucretius, this was taken up by Leonardo da Vinci, Galileo and, more recently, by Primo Levi and Italo Calvino. Of these, da Vinci probably did not know Lucretius' works but the others revered him. The majesty of Galileo's writing surely owes something to Lucretius:

For my part I consider the earth very noble and admirable precisely because of the diverse alterations, changes, generations, etc. that occur in it incessantly. If not being subject to any changes, it were a vast desert of sand or a mountain of jasper, or if at the time of the flood the waters which covered it had frozen, and it had remained an enormous globe of ice where nothing was ever born or ever altered or changed, I should deem it a useless lump in the universe, devoid of activity and, in a word, superfluous and essentially nonexistent.
From *Dialogue Concerning the Two Chief World Systems* (translator: Stillman Drake) 1632

'Poetry and science' at first suggests poems about science, but it can also help us to look at science from a different angle. Reading a scientific paper that opens up a new field has a similar effect to reading a significant new voice in poetry for the first time: you are not sure where it will lead.

Take, for example, the paper by Joanna Aizenberg and colleagues which revealed the single-crystal microlens array in the brittle-star *Ophiocoma wendtii*¹. The structure of the array itself is beautiful: it resembles the sort of organic blob architecture that is now fashionable. Each lens is microengineered to an astonishing degree: it corrects the spherical aberration of light by means of the classic Huygens/Descartes reversed-curve rear surface, and the precise alignment of the crystal excludes double-reflection effects. Then there was the synthetic DNA octahedron that featured on the cover of *Nature* in February 2004 (ref. 2). I claim these finds as poetry.

Poetry is one way of trying to reconcile different and contradictory aspects of life. Poetry is another viewpoint, a way of getting stereoscopic vision or at least perspective. James Watson is not a poet, but when he

wrote an account of his work on the double helix he gave a different perspective compared with that given by the famous paper published in *Nature*. People were shocked or delighted because it told a human story. It told of science, but in the voice of someone who also chases girls, is competitive and does the million-and-one things we all do. A scientific culture that eschewed such an approach completely and insisted that only the academic papers mattered would not be a culture: it would be pedantic formalism gone mad.

I believe that science is more involved, more high-pressure than poetry. The Nobel Prize committees do not have to scrape any barrels to find great scientific discoveries each year. Really significant poetry emerges much less often than impressive scientific discoveries. But not all scientific papers are Earth-shattering either: most of the time most poetry magazines and most scientific journals contain bread-and-butter work.

It is interesting to look at the very few figures who were equally strong scientists and poets. In *The Periodic Table* Primo Levi wrote of his scientific awakening at school:

Everything around us was a mystery pressing to be revealed: the old wood of the benches, the sun's sphere beyond the window panes and the roofs, the vain flight of the pappus down in the June air.

In this passage, you can see a man who might become a poet or a scientist or both. In Levi's case his life was irretrievably skewed by the War and his deportation to Auschwitz. Surviving and returning to the chaos of post-war Italy, Levi renounced any hopes of a career in research and instead became a skillful industrial chemist and ultimately a great writer. But he kept up an interest in pure science, especially biology, and wrote many beautiful short essays on it.

Anyone who feels the passion Levi felt as a boy has three choices: to be a scientist; to write; or to paint. These are three ways of apprehending the material world and it is rare to find all three modes in one person. Da Vinci is the most perfect and perhaps the only example. He would take the same phenomenon and try to exhaust its possibilities in all three ways. He studied the turbulent motion of water because he needed to understand how to paint this phenomenon as part

of one of his compositions. In his sketchbooks he tried to formalize the problem by drawing geometrical curlicues. (This was da Vinci the frustrated scientist working before the mathematical formalism was introduced by Galileo and Newton.) And finally, he tried to describe it in words. Here is his famous deluge — da Vinci trying to be Milton before his time:

But the swollen waters should be coursing round the pool which confines them, and striking against various obstacles with whirling eddies, leaping up into the air in turbid foam, and then falling back and causing the water where they strike to be dashed up into the air; and the circling waves which recede from the point of contact are impelled by their impetus right across the course of the other circling waves which move in an opposite direction to them, and after striking against these they leap up into the air without becoming detached from their base. And where the water issues forth from the said pool, the spent waves are seen spreading out towards the outlet; after which, falling or descending through the air, this water acquires weight and impetus; and then piercing the water where it strikes, it tears it apart and dives down in fury to reach its depth, and then recoiling it springs back again towards the surface of the lake accompanied by the air which has been submerged with it...

All three modes obviously have their strengths and weaknesses. It might be a waste of time trying to express quantum theory in words. And it is pointless trying to mathematicize the details of every dab of paint in a painting. But people often pursue one route while being blind to the others: a Leonardo da Vinci will always be rare, even at the level of intention, let alone achievement.

Although poetry has largely withdrawn from miltonic grandeur to occupy a place somewhere between the adverts on underground trains, the idea of the poetic — something sublime, intensely moving, grand or awesome — retains its power. It is sheer poetry we say of an Andrew Flintoff six, a beautifully designed car, Concorde or an elegant scientific experiment. But I think it is fair to say that, awesome though it is, a great deal of what constitutes science is

resistant to poetry. Some of its more resonant conclusions attract poets, but the processes of science are too nested to suit poetic narrative.

Take a classic experiment such as the Meselson–Stahl experiment, which showed that in DNA replication, the two helices of DNA separate, and a complementary strand is added to each helix. In 1958, it was not known whether DNA replicated in this way or by keeping the double helix intact, with new complementary chains being added on the outside. The latter scenario was much less likely but how could you prove the first mechanism was correct? The conclusive experiment is very typical of science, involving a logical series of steps to falsify hypotheses.

DNA is now an icon of our age: there are poems about it. But the real science behind it, of which the Meselson–Stahl experiment is typical, does not feature in poems, nor will it ever do so, unless some bloody-minded poet takes up the challenge to prove me wrong.

So what do poets write about when they write about science? Mimicry is inherently close to poetry: the crux being simile, that word ‘like’. In the early 1980s, a school of poetry that relied almost entirely on simile and metaphor for its effects briefly flourished: poetry seemed to become a pageant of mimicry.

*Seven striped shallots are giving
A silent recital on their sitars*

*And the magnolia dances motionless,
Like a deity with many arms...*
From *Shallots* by Craig Raine

In its fantasciation and wilful image-making some will find this the antithesis of science. But nature herself is a ludic poet: the elaborate appearance of many creatures seems far in excess of anything dictated by mere survival.

If simile is a key poetic strategy for embracing the physical world it is often big and little things that are compared. The novelist Vladimir Nabokov said:

*There is, it would seem, in the dimensional
scale of the world a kind of delicate
meeting place between imagination and
knowledge, a point arrived at by
diminishing large things and enlarging*

small ones, that is intrinsically artistic.
From *Speak, Memory*

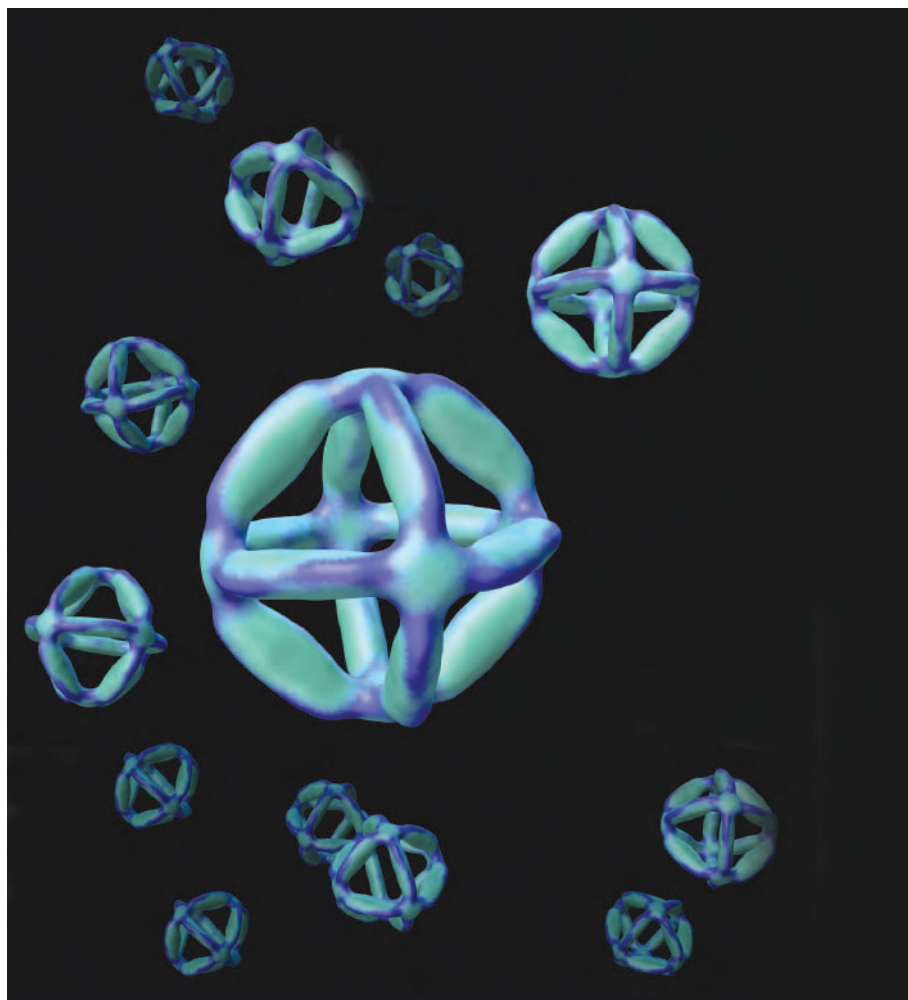
And this is what poets tend to do when they write about science: they make big things (such as the Universe) small, and small things big:

*Curled like the scrolled end of a chair's
arm, swaddled
In dust or gas, the comet's rock or ice
heart
Went to bits in the field
Of invisible stress round Jupiter...*
From *The Jupiter Collisions* by Lachlan Mackinnon

*Whole eras, seemingly without event,
Now scud the glassy pool processionally*

*Until one day, misty, uncalendered,
As mild and emphatic as a schwa,
Vascular tissue, conduit filaments
Learn how to feed the outpost of that
small
Emerald principate.*
From *Green: An Epistle* by
Anthony Hecht

But there are other ways of bringing science to poetry. Miroslav Holub eschewed this kind of micro-precise description. He seized on the starkness of science, and this was no doubt aided by three factors: he spent most of his life under a communist regime; he was a Czech and there is streak of deadpan drollery in Czechs; and, most importantly, he was as much a scientist as a poet, being a



Poetic discovery: DNA assembled into nanoscale octahedra.

M. PIQUE

noted immunologist. Holub did not soften science with poetry; he imported scientific rigour into life.

*To amuse His Royal Majesty he will
change water into wine.*

....
*Then along comes a student and asks:
Think up sine alpha
Greater than one.*

*And Zito grows pale and sad: terribly
sorry. Sine is
Between one and minus one. Nothing you
can do about that.
From Zito the Magician*

Science and poetry are serious subjects and to yoke them together is to be even more portentous, but we must not get too po-faced about it. W. H. Auden insisted that “The pious fable and the dirty story / share in the

total literary glory”. Science and poetry often meet in humour. It would be churlish not to recognize that there is a tradition of scientists — sometimes great scientists — who write poetry. It is usually light verse and it tends to the squib: the paradoxical nature of much of science lends itself to the crisp closure of tight rhymes: “There was a young lady called Bright / who travelled much faster than light. / She went out one day / in a relative way / and came back the previous night”. At a slightly higher level, James Clerk Maxwell wrote verse such as this:

*So, down through untold generations,
transmission of structureless germs
Enables our race to inherit the thoughts of
beasts, fishes and worms.*

James Fenton is a poet who might have been a scientist. In his early days he would scour the Pitt Rivers Museum in Oxford and find poetry in obscure mycological papers: “Wallroth / described the fungus as possessing / a pezizaform hairy sporochodium / with a flattened powdery surface layer / Of globose, vesiculate, hyaline conidia”.

The most frivolous but irresistible kind of found poetry in science belongs to the Cognomen Syndrome — Tom Stoppard’s mock-scientific name for people whose name matches their job. The German aerodynamicist Dietrich Bechert pointed out to me that many German scientists active in the field of biomimetics, or Bionik as it is known in Germany, have Stoppardian names: Nachtigall (nightingale), Hummel (bumble bee), Spätz (sparrow). The man in charge of an institute for bees is called Bienefeld (bee field). Then there is Professor Fisch who works on fish...

Nature recently furnished a perfect example of the Cognomen Syndrome³: “On the straight and narrow” by Richard A. Register. “Nanoscale chemical patterns written on a substrate can direct the self-assembly of polymer overlays with remarkable precision.”

immortalibu’ de diuis dare dicta suerit
atque omnem rerum naturam pandere dictis.

The world is too full of flaws to be divinely created

quod si iam rerum ignorem primordia quae sint
hoc tamen ex ipsis caeli rationibus ausim
confirmare aliisque ex rebus reddere multis,
nequaquam nobis diuinitus esse paratam
naturam rerum: tanta stat praedita culpa.
principio quantum caeli tegit impetus ingens,
de auidam partem montes silvaeque ferarum
re, tenent rupes uastaeque paludes
od late terrarum distinet oras.
prope partis fervidus ardor
casus mortalibus aufert.
amen id natura sua ui
humana resistat,
anti

Could it be that a man with the name Register is drawn, in his career, towards the precise alignment of overlays: Register his name and registration his trade?

Science and poetry: two great things. So why mix them up? Because life is short and to know only one thing is sad. But it is worth remembering Samuel Johnson’s stern admonition: “The truth is that the knowledge of external nature and of the sciences which that knowledge requires or includes, is not the great or frequent business of the human mind...we are all moralists but we are geometricians only by chance. Our intercourse with intellectual nature is necessary; our speculations upon matter are voluntary, and at leisure”. So for Johnson the proper subject of poetry lies in this human moral universe. He used to quote with approval, “the poet does not number the streaks on the tulip”. But there will always be poets who are drawn to those tulips and some tulip-streak enumerators who hunger for the wider picture. ■

Peter Forbes is a science writer and former editor of Poetry Review. The Gecko’s Foot: Engineering New Materials from Nature will be published by Fourth Estate in August.

1. Aizenberg, J., Tkachenko, A., Weiner, S., Addadi, L. & Hendler, G. *Nature* **412**, 819–822 (2001).
2. Shih, W. M., Quispe, J. D. & Joyce, G. F. *Nature* **427**, 618–621 (2004).
3. Register, R. A. *Nature* **424**, 378–379 (2003).

Miroslav Holub used poetry to encapsulate the starkness of science.

The DNA sequence of the human X chromosome

A list of authors and their affiliations appears at the end of the paper

The human X chromosome has a unique biology that was shaped by its evolution as the sex chromosome shared by males and females. We have determined 99.3% of the euchromatic sequence of the X chromosome. Our analysis illustrates the autosomal origin of the mammalian sex chromosomes, the stepwise process that led to the progressive loss of recombination between X and Y, and the extent of subsequent degradation of the Y chromosome. LINE1 repeat elements cover one-third of the X chromosome, with a distribution that is consistent with their proposed role as way stations in the process of X-chromosome inactivation. We found 1,098 genes in the sequence, of which 99 encode proteins expressed in testis and in various tumour types. A disproportionately high number of mendelian diseases are documented for the X chromosome. Of this number, 168 have been explained by mutations in 113 X-linked genes, which in many cases were characterized with the aid of the DNA sequence.

The X chromosome has many features that are unique in the human genome. Females inherit an X chromosome from each parent, but males inherit a single, maternal X chromosome. Gene expression on one of the female X chromosomes is silenced early in development by the process of X-chromosome inactivation (XCI), and this chromosome remains inactive in somatic tissues thereafter. In the female germ line, the inactive chromosome is reactivated and undergoes meiotic recombination with the second X chromosome. The male X chromosome fails to recombine along virtually its entire length during meiosis: instead, recombination is restricted to short regions at the tips of the X chromosome arms that recombine with equivalent segments on the Y chromosome. Genes inside these regions are shared between the sex chromosomes, and their behaviour is therefore described as 'pseudoautosomal'. Genes outside these regions of the X chromosome are strictly X-linked, and the vast majority are present in a single copy in the male genome.

The unique properties of the X chromosome are a consequence of the evolution of sex chromosomes in mammals. The sex chromosomes have evolved from a pair of autosomes within the last 300 million years (Myr)¹. In the process, the original, functional elements have been conserved on the X chromosome, but the Y chromosome has lost almost all traces of the ancestral autosome, including the genes that were once shared with the X chromosome. The hemizyosity of males for almost all X chromosome genes exposes recessive phenotypes, thus accounting for the large number of diseases that have been associated with the X chromosome². The characteristic pattern of X-linked inheritance (affected males and no male-to-male transmission) was recognized by the eighteenth century for some cases of haemophilia, and gave impetus in the 1980s to the earliest successes in positional cloning—of the genes for chronic granulomatous disease³ and Duchenne muscular dystrophy⁴. For females, the major consequence of the loss of genes from the Y chromosome is XCI, which equalizes the dosage of X-linked gene products between the sexes.

The biological consequences of sex chromosome evolution account for the intense interest in the human X chromosome in recent decades. However, evolutionary processes are likely to have shaped the behaviour and structure of the X chromosome in many other ways, influencing features such as repeat content, mutation rate, gene content and haplotype structure. The availability of the finished sequence of the human X chromosome, described here, now allows us to explore its evolution and unique properties at a new level.

The X chromosome sequence

We constructed a map of the X chromosome using predominantly

P1-artificial chromosome (PAC) and bacterial artificial chromosome (BAC) clones (Supplementary Table 1), which were assembled into contigs using restriction-enzyme fingerprinting and integrated with earlier maps using sequence-tagged site (STS) content analysis⁵. Gaps were closed by targeted screening of clone libraries in bacteria or yeast, and by assessing BAC and fosmid end-sequence data for evidence of spanning clones. Fourteen euchromatic gaps remain intractable, despite using libraries with a combined 80-fold chromosome coverage. Five of these gaps are within the 2.7 megabase (Mb) pseudoautosomal region at the tip of the chromosome short arm (PAR1). This is reminiscent of the situation in other human sub-telomeric regions⁶, and might reflect cloning difficulties in an area with a high content of (G+C) nucleotides and mini-satellite repeats.

We selected 1,832 clones from the map for shotgun sequencing and directed finishing using established procedures⁷. Finished sequences were estimated to be more than 99.99% accurate by independent assessment⁸. The sequence of the X chromosome has been assembled from the individual clone sequences and comprises 16 contigs. These extend into the telomeric (TTAGGG)_n repeat arrays at the ends of the chromosome arms, and include both pseudoautosomal regions (PARs). The data were frozen for the analyses described below, at which point we had determined 150,396,262 base pairs (bp) of sequence (Supplementary Table 2). Subsequently, we obtained a further 609,664 bp of sequence. The 14 euchromatic gaps are estimated to have a combined size of less than 1 Mb (see Methods and Supplementary Table 2), and the sequence therefore covers at least 99.3% of the X chromosome euchromatin. There is also a single heterochromatic gap corresponding to the polymorphic 3.0 (±0.4) Mb array⁹ of alpha satellite DNA at the centromere. On this basis, we conclude that the X chromosome is approximately 155 Mb in length.

The coverage and quality of the finished sequence have been assessed using independent data. All markers from the deCODE genetic map¹⁰ are found in the sequence and the concordance of marker orders is excellent with only one discrepancy. DXS6807 is the most distal Xp marker on the deCODE map (4.39 cM), but in the sequence this marker is proximal to three others with genetic locations of 9–11 cM on the deCODE map. Out of 788 X-chromosomal RefSeq¹¹ messenger RNAs that were assessed, 783 were found completely in the sequence, and parts of four others are also present (T. Furey, personal communication). The missing segments of *GTPBP6*, *CRLF2*, *DHRX* and *FGF16* lie within gaps 1, 4, 5 and 10, respectively, and the *GAGE3* gene is in gap 7 (Supplementary Table 2). The sequence assembly was assessed using fosmid end-sequence pairs that match the X chromosome sequence. The

orientation and separation of end-pairs of more than 17,000 fosmids were consistent with the sequence assembly. In two cases, sequences had been misassembled owing to long and highly similar repeats. There were six instances of large deletions in sequenced clones, which were resolved by determining fosmid sequences through the deleted regions. Finally, there were two cases of apparent length variation between the reference sequence and the DNA used for the fosmid library.

Features of the X chromosome sequence

The annotated sequence of the X chromosome is presented in Supplementary Fig. 1, and updates are contained in the Vertebrate Genome Annotation (VEGA) database (http://vega.sanger.ac.uk/Homo_sapiens/). The distribution of a number of sequence features on the chromosome is shown in Fig. 1. Analysis of the sequence reveals a gene-poor chromosome that is highly enriched in interspersed repeats and has a low (G+C) content (39%) compared with the genome average (41%).

Genes

Based on a manual assessment of all publicly available human expressed sequences and genes from other organisms, we have annotated 1,098 genes (7.1 genes per Mb) across four different categories (see Methods): known genes (699), novel coding sequences (132), novel transcripts (166), and putative transcripts (101). We have also identified 700 pseudogenes in the sequence (4.6 pseudogenes per Mb), of which 644 are classified as processed and 56 as non-processed. The gene density (excluding pseudogenes) on the X chromosome is among the lowest for the chromosomes that have been annotated to date. This might simply reflect a low gene density on the ancestral autosomes. Alternatively, selection may have favoured transposition of particular classes of gene from the X chromosome to the autosomes during mammalian evolution. These could include developmental genes for which the protein products are required in double dose in males (or in females after XCI has occurred), or genes for which mutation in male somatic tissues is lethal.

Physical characteristics of the genes and pseudogenes are summarized in Supplementary Table 3. Exons of the 1,098 genes account for only 1.7% of the X chromosome sequence. On the basis of the lengths of these gene loci, 33% of the chromosome is transcribed. This is considerably below the recent estimates for chromosomes 6 (ref. 12), 9 (ref. 6), 10 (ref. 13) and 13 (ref. 14), to which the equivalent gene annotation procedure was applied (Supplementary Table 4), and is a reflection not just of low gene density on chromosome X but also of low gene length. For example, mean gene length is 49 kilobases (kb) on chromosome X compared with 57 kb on chromosome 13. Nevertheless, the X chromosome contains the largest known gene in the human genome, the dystrophin (*DMD*) locus in Xp21.1, which spans 2,220,223 bp. Consistent with its low gene density, the frequency of predicted CpG islands on the X chromosome is only 5.25 per Mb, which is exactly half of the estimated genome average⁷. There is an association with a CpG island for 49% of the known genes, the category for which the most complete gene structures are expected in the current annotation.

We identified evolutionarily conserved regions (ECRs) by comparing the X chromosome sequence to the genomes of mouse, rat, zebrafish and the pufferfishes *Tetraodon nigroviridis* and *Fugu rubripes* (Supplementary Table 5). There are 4,493 ECRs that are conserved between the X chromosome and all of the other species. Of these, 4,393 overlap with 4,373 annotated exons. The remaining 100 ECRs are most likely to be unannotated exons, although some could be highly conserved control or structural elements. From these data we conclude that we have annotated at least 97.8% of the protein-coding exons on the X chromosome $[(4,373/(4,373+100)) \times 100]$.

Non-coding RNA genes

The gene set described above includes non-coding RNA (ncRNA) genes only when there is supporting evidence of expression from complementary DNA or expressed-sequence-tag (EST) sources. Using a complementary approach, we analysed the X chromosome sequence using the Rfam¹⁵ database of structural RNA alignments, and predicted 173 ncRNA genes and/or pseudogenes (Supplementary Fig. 1 and Supplementary Table 6). These are physically separate from the genes described in the preceding section and are not included in the total gene count, owing to the difficulty in discriminating between genes and pseudogenes for these ncRNA predictions. Using tRNAscan-SE¹⁶, we predicted only two transfer RNA genes on the X chromosome (Supplementary Table 6), out of the several hundred predicted in the human genome⁷. Thirteen microRNAs from the microRNA registry¹⁷ have also been mapped onto the sequence (Supplementary Table 7).

The most prominent of the ncRNA genes on the X chromosome is *XIST* (X (inactive)-specific transcript)¹⁸, which is critical for XCI. The *XIST* locus spans 32,103 bp in Xq13, and its untranslated transcript coats and transcriptionally silences one X chromosome in *cis*. The RefSeq¹¹ transcript of *XIST* is an RNA of 19,275 bases, which includes the largest exon on the chromosome (exon 1: 11,372 bp). There is also evidence for shorter *XIST* transcripts generated by alternative splicing, particularly in the 3' region of the gene¹⁹. In the mouse, *Tsix* is antisense to *Xist*²⁰, and its transcript (or the process of its transcription) is believed to repress the accumulation of *Xist* RNA. There is evidence for transcription antisense to *XIST* in human^{21,22}, but we have been unable to annotate the human *TSIX* gene as there are no corresponding expressed sequences in the public databases, and because there is a lack of primary sequence conservation between the human and mouse regions. In the human sequence, two other ncRNA genes are annotated in the 400 kb region distal to *XIST*, which are orthologues of the mouse genes described previously as *Jpx* and *Ftx* (ref. 23). In the mouse, *Xist*, *Jpx* and *Ftx* are located within a smaller area of approximately 200 kb²³.

The cancer-testis antigen genes

On assessing the predicted proteome of the X chromosome for Pfam²⁴ domains, our most prominent finding was the presence of the MAGE domain (IPR002190) in 32 genes (Supplementary Table 8). In comparison, only four other MAGE genes are reported in the rest of the genome: *MAGEF1* on chromosome 3, and *MAGEL2*, *NDN* and *NDNL2* on chromosome 15. The MAGE gene products are members of the cancer-testis (CT) antigen group, which are characterized by their expression in a number of cancer

Figure 1 Features of the X chromosome sequence. **a**, X chromosome ideogram according to Francke⁶⁵. **b**, Evolutionary domains of the X chromosome: the X-added region (XAR), the X-conserved region (XCR; dotted region in proximal Xp does not appear to be part of the XCR), the pseudoautosomal region PAR1, and evolutionary strata S5–S1. **c**, Sequence scale in intervals of 1 Mb. Note that correlation between cytogenetic band positions and physical distance is imprecise, owing to varying levels of condensation of different Giemsa bands. **d**, (G+C) content of 100-kb sequence windows. **e**, Number of genes in 1-Mb sequence windows (pseudogenes not included). **f–k**, Fractional coverage of 100-kb sequence windows by interspersed repeats: L1 repeats (**f**), L1M subfamilies of L1 repeats (**g**), L1P subfamilies of L1 repeats (**h**), *Alu* repeats (**i**), L2 repeats (**j**), MIR repeats (**k**). Vertical grey lines in **d–k** represent gaps in the euchromatic sequence of the chromosome. Grey bar centred at approximately 60 Mb shows the position of the centromere. **l**, A selection of landmark genes on the chromosome. *OPN* refers to the three opsin genes in the reference sequence, which are organized as follows: *cen-OPN1LW-OPN1MW-OPN1MW-tel*. **m**, Genes that escape from X-chromosome inactivation as previously identified⁴⁸. **n**, Cancer-testis antigen genes, belonging to the MAGE (light green), GAGE (dark green), SSX (magenta), SPANX (orange) or other (grey) CT gene families. For the genes in **l–n**, arrows indicate the direction of transcription.

types, while their expression in normal tissues is solely or predominantly in testis. This expression profile has led to the suggestion that the CT antigens are potential targets for tumour immunotherapy. A recent report listed 84 CT antigen genes for the human genome²⁵. The X chromosome gene set we describe above contains 99 CT antigen genes and includes novel members of the *MAGE*, *GAGE*, *SSX*, *LAGE*, *CSAGE* and *NXF* families (Supplementary Table 9). Assessment of the most recent RefSeq¹¹ information shows that this set does not include two known *MAGE* genes (*MAGEA5* and *MAGEA7*) and seven *GAGE* genes (*GAGE3–7*, *7B* and *8*), which are expected to lie in gaps 14 and 7, respectively (Supplementary Table 2). Furthermore, gaps 6 and 9 are also within regions of CT antigen gene duplication. Therefore, we predict that approximately 10% of the genes on the X chromosome are of the CT antigen type.

Conclusive data on the normal functions of the CT antigens, or their involvement in disease conditions, are very limited. However, the remarkable enrichment for CT antigen genes on the X chromosome relative to the rest of the genome might be indicative of a male advantage associated with these genes. Recessive alleles that are beneficial to males are expected to become fixed more rapidly on the X chromosome than on an autosome²⁶. If these alleles are detrimental to females, their expression could become restricted to male tissues as they rise to fixation. Both the concentration of the CT antigen genes on the X chromosome and their expression profiles are consistent with this model of male benefit. The CT antigen genes on the X chromosome are also notable for the expansion of various gene families by duplication. This degree of duplication is perhaps an indication of selection in males for increased copy number. In

this context, it is of interest that the *MAGE* family has independently expanded on the X chromosome in both the human and mouse lineages²⁷.

Repetitive sequences

Interspersed repeats account for 56% of the euchromatic X chromosome sequence, compared with a genome average of 45% (Supplementary Table 10). Within this, the *Alu* family of short interspersed nuclear elements (SINEs) is below average, in keeping with the gene-poor nature of the chromosome. Conversely, long terminal repeat (LTR) retroposon coverage is above average; but the most remarkable enrichment is for long interspersed nuclear elements (LINEs) of the L1 family, which account for 29% of the X chromosome sequence compared to a genome average of only 17%. The possible significance of this enrichment for XCI is discussed later.

Applying the criterion of at least 90% sequence identity over at least 5 kb (ref. 28), we estimate that intrachromosomal segmental duplications account for 2.59% of the X chromosome (Supplementary Table 11 and Supplementary Fig. 2). In contrast, interchromosomal segmental duplications indicated by sequence matches to the autosomes account for a very small fraction (0.24%) of the X chromosome (Supplementary Table 12). Six gaps in the X chromosome map are either flanked by or contained within intrachromosomally-duplicated segments (gaps 2, 3, 6, 7, 9 and 14 in Supplementary Table 2), which might produce instability of clones or otherwise confound mapping progress. The intrachromosomal duplicates are striking in their proximity. Apart from the two

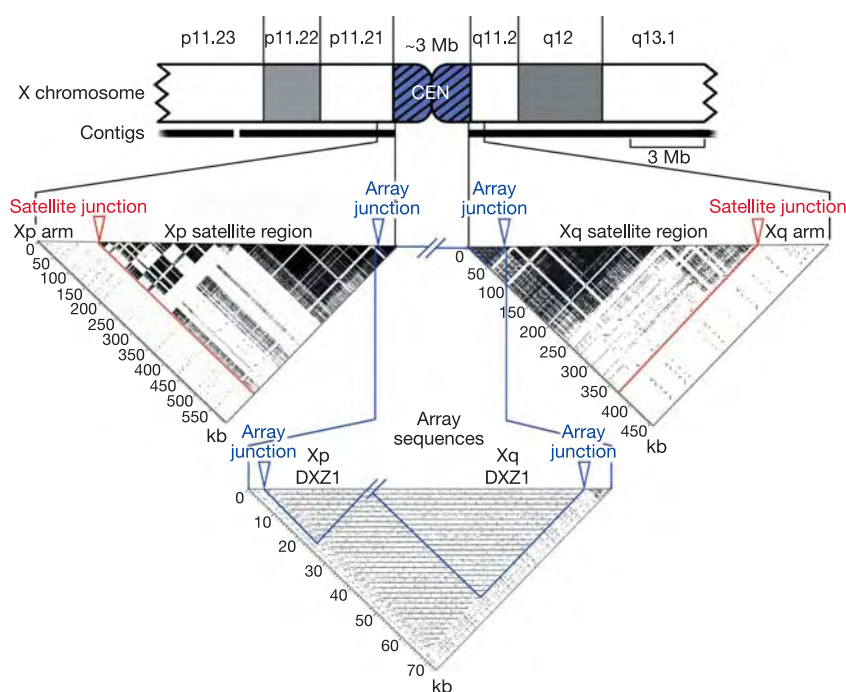


Figure 2 Xp and Xq pericentromeric contigs extend into the X-chromosome-specific higher-order alpha satellite, DXZ1. The pericentromeric region of the X chromosome is shown as a truncated ideogram. Self-self alignments of proximal sequences from each arm are illustrated by dot plots below the ideogram. On each plot, the junction between the arm sequence and the arm-specific satellite region is marked by a red arrow, and the junction between the arm-specific satellite region and the X-chromosome-specific alpha satellite array (DXZ1) is marked with a blue arrow. Approximately 594 kb of sequence were analysed from Xp, including ~21 kb of DXZ1 sequence. The ~454 kb of sequence analysed from Xq included ~44 kb of DXZ1 sequence. In each case, ~100 kb of arm sequence were included. The highly repetitive structure of pericentromeric satellites is in stark contrast to the near absence of repetitive structure in the arm sequences, despite an

unusually high density of LINE repeats in these regions. Gaps in the dark satellite regions occur where interspersed elements (LINEs, SINEs and LTRs) interrupt the satellite sequences. In the Array Sequences dot plot, the most proximal ~21 kb of the Xp sequence is joined to the most proximal ~44 kb of the Xq sequence. The periodic nature of the centromeric, higher-order alpha satellite array is evident. Black horizontal lines on the plot reveal near identity of sequences spaced at ~2 kb intervals. This DXZ1 sample represents ~65 kb of the 3 (± 0.4) Mb alpha satellite array. The regions outlined in blue are self-self alignments ('Xp DXZ1' and 'Xq DXZ1'), and the remaining rectangular region of the plot is an alignment of Xp versus Xq DXZ1, which reveals the close relationship between DXZ1 sequences from each arm.

segments containing *SSX* gene copies, which are separated by 4.5 Mb, only six of 229 matches are separated by more than 1 Mb. Among these duplications are well-described cases that are associated with genomic disorders²⁹. In Xp22.32, deletions of the steroid sulphatase (*STS*) gene, causing X-linked ichthyosis (Online Mendelian Inheritance in Man (OMIM)² entry number 308100), result from recombination between flanking duplications that contain copies of the *VCX* gene. Also, some instances of Hunter syndrome (OMIM 309900), red-green colour blindness (OMIM 303800), Emery-Dreifuss muscular dystrophy (OMIM 310300), incontinentia pigmenti (OMIM 308300) and haemophilia A (OMIM 306700) result from rearrangements involving duplicated sequences in Xq28. In haemophilia A, mutations are frequently the result of inversions between a sequence in intron 22 of the *F8* gene and one of two more distally located copies. A novel finding from our analysis of the X chromosome reference sequence is that the two distal copies are in opposite orientations. Therefore, a large deletion involving *F8* and several more distal genes could be an alternative to the inversion rearrangement. A deletion consistent with this prediction has been reported in a family in which carrier females are affected by a high spontaneous-abortion rate in pregnancy³⁰.

The X chromosome centromere

The X chromosome sequence extends from both arms into centromeric, higher-order repeat sequences, which are known to be associated functionally with the X centromere^{31–33}. The most proximal 494 kb and 360 kb of the Xp and Xq sequences, respectively, consist of extensive regions of satellite DNA, adjacent to

euchromatin of the chromosome arms that is exceptionally high in L1 content (Fig. 2). The satellite region on Xp contains small amounts of other satellite families³¹, whereas that on Xq consists entirely of alpha satellite. Similar to all other human chromosome arms that have been examined^{33,34}, these transition regions consist of monomeric alpha satellite that is not associated with centromere function. Both the Xp and Xq contigs reported here, though, extend more proximally and reach into highly homogeneous, higher-order repeat alpha satellite (DXZ1). Critically, the Xp and Xq contig copies of the DXZ1 repeat are themselves 98–100% identical in sequence, and are oriented in the same direction along the chromosome (Fig. 2). On this basis, the two contigs reach the 'end' of each chromosome arm and thus also reach the centromeric locus from either side. This represents a logical endpoint for efforts to complete the sequence of chromosome arms in the human genome, and the first demonstration of this endpoint is provided by the X chromosome sequence.

Single-nucleotide polymorphisms

A total of 153,146 candidate single-nucleotide polymorphisms (SNPs) have been mapped onto the X chromosome sequence and are displayed in the VEGA database. These include 901 SNPs that result in non-synonymous changes in protein-coding regions, and are therefore candidate functional protein variants. The heterozygosity level on the X chromosome is known to be well below that of the autosomes, and this difference can be explained partly or entirely by population genetic factors³⁵. Included in the mapped SNPs are 62,334 that were identified by alignment of flow-sorted

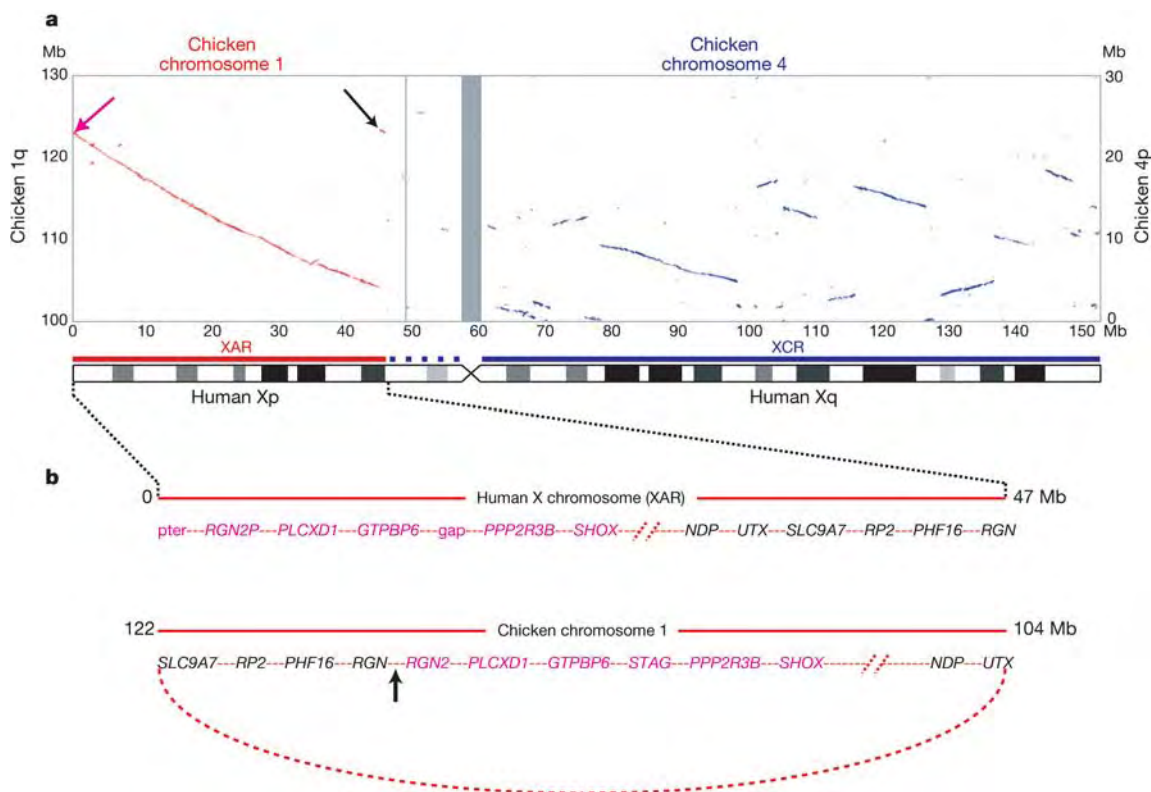


Figure 3 Homologies between the human X chromosome and chicken autosomes. **a**, Plot of BLASTZ sequence alignments between the X chromosome and chicken chromosomes 1 (red) and 4 (blue). Grey bar centred at approximately 60 Mb shows the position of the X centromere. Only the relevant section of each chicken chromosome is shown (see Mb scale at left for chromosome 1 and at right for chromosome 4). A schematic interpretation of the homologies shows the XAR and XCR as red and blue bars, respectively (see Fig. 1). Homologies at the ends of the XAR are indicated with arrows and are expanded in **b**.

b, (Top) Genes at the ends of the human XAR. Genes from distal Xp (magenta arrow in **a**) are in magenta and genes from Xp11.3 (black arrow in **a**) are shown in black. (Bottom) Arrangement of the orthologous genes on chicken chromosome 1. A hypothetical ring chromosome, with the equivalent gene order to that observed in the chicken, is indicated by the curved, dotted red line. Recombination between one end of the established X chromosome and the ring chromosome at the arrowed position could, in a single step, have added the XAR and created the gene order observed on the human X chromosome.

X chromosome shotgun sequence reads to the X chromosome reference sequence. Using comparable sequence data for chromosome 20, we calculated that the heterozygosity level on the X chromosome is approximately 57% of that observed for the autosome.

Evolution of the human X chromosome

Males of the three mammalian groups—Eutheria ('placental' mammals), Metatheria (marsupials) and Prototheria (egg-laying mammals)—have X and Y sex chromosomes. Ohno proposed in 1967 that the mammalian sex chromosomes evolved from an autosome pair following their recruitment into a chromosomal system for sex determination¹. A barrier to recombination developed between these 'proto' sex chromosomes, isolating the sex-determining regions and eventually spreading throughout the two homologues. In the absence of recombination, the accumulation of mutation events subsequently led to the degeneration of the Y chromosome. The sex chromosomes of birds are not homologous to those of the mammals. The sex chromosome system of birds evolved independently during the last 300 Myr, giving rise to homogametic (ZZ) male birds and heterogametic (ZW) female birds, in contrast to the mammalian system of XY males and XX females.

The autosomal origin of the mammalian sex chromosomes is vividly illustrated by alignment of the human X and chicken whole genome sequences (Fig. 3a). Orthologues of some human X chromosome genes were previously mapped to chicken chromosomes 1q13-q21 and 4p11-p14 (ref. 36). Using genomic sequence alignment, we identified approximately 30 regions of homology that together cover most of human Xq and are confined to a single section of approximately 20 Mb at the end of chicken chromosome 4p (Fig. 3a). In contrast, most of the short arm (Xp11.3-pter), including the pseudoautosomal region PAR1, matches a single block of chicken chromosome 1q. No clear picture emerges regarding the origin of the remainder of the short arm (Xcen-p11.3). We were unable to detect large regions of conserved synteny using sequence alignment, and genes from this region have orthologues on several chicken autosomes, including chromosomes 12, 1 and 4 (ref. 37). This region is also characterized by the expansion of several families of CT antigen genes (Fig. 1), which have no readily detectable orthologues in chicken. The present analysis supports the notion of a mammalian 'X-conserved region' (XCR)³⁸, which includes the long arm and is descended from the proto-X chromosome. It also supports a separate, large addition ('X-added region' or XAR³⁸) to the established X chromosome by translocation from a second autosome, which occurred in the eutherian mammals before their radiation (~105 Myr ago). In contrast to earlier hypotheses, however, it appears that much of the proximal short arm (Xcen-p11.3) should no longer be considered part of an XCR.

The precise location of genes that demarcate the XAR suggests a possible mechanism for the addition. The annotated genes at the extreme ends of the 47 Mb XAR are *PLCXD1* (cU136G2.1 in Supplementary Fig. 1) near Xpter, and *RGN* in Xp11.3. We also found an unprocessed *RGN* pseudogene (*RGN2P*) at Xpter, distal to *PLCXD1*. The orthologues for these three loci are adjacent on chicken chromosome 1, in the order (tel)—*RGN*—*RGN2*—*PLCXD1*—(cen) (Fig. 3b). The generation of these two different gene orders from a common ancestral sequence would require a minimum of two rearrangements as well as the translocation that added the XAR. A more parsimonious model suggested by these data, however, is that the XAR was acquired by recombination between the X chromosome and a ring chromosome in which the ancestral *PLCXD1*, *RGN* and *RGN2* sequences were neighbours (Fig. 3b).

In order to examine more recent patterns of evolution, we compared the human X chromosome with other mammalian sequences. We saw nine major blocks of sequence homology between human and mouse X chromosomes, and eleven between human and rat (Fig. 4). The homology blocks occupy almost the

entirety of each X chromosome, confirming the remarkable degree of conserved synteny of this chromosome within the eutherian mammalian lineage. This is consistent with Ohno's law, which predicts that the establishment of a dosage compensation mechanism had a stabilizing effect on the gene content of the mammalian X chromosome¹. On the long arm, just two blocks of homology account for the entire alignment of the human and corresponding mouse sequences, but the mouse homologous regions are punctuated with three additional segments, each containing long and very similar repeats (arrowed in Fig. 4). Alignment of human Xq with the rat sequence reveals four discrete homology blocks; the greater fragmentation compared with the mouse alignment would be explained by a minimum of two rearrangements, one in each of the two mouse-human homology blocks, specifically on the rat lineage. The mouse-specific repeat segments are not detected in the current version of the rat genome sequence. On the short arm of the human X chromosome, seven major blocks of homology with each rodent account for most of the human sequence (Fig. 4). Using the dog as an outgroup, we established that the human and dog X chromosome sequences are essentially collinear (K. Lindblad-Toh, personal communication). Therefore, all of the rearrangements indicated in Fig. 4 occurred in the rodent lineage, and the human X chromosome appears to have been remarkably stable in its organization since the radiation of eutherian mammals. This is consistent with the recent prediction, derived from a comparison of human, rodent and chicken chromosomes, that the human X chromosome is identical to the putative ancestral (eutherian) mammalian X chromosome³⁹.

The most notable difference we found between the human and

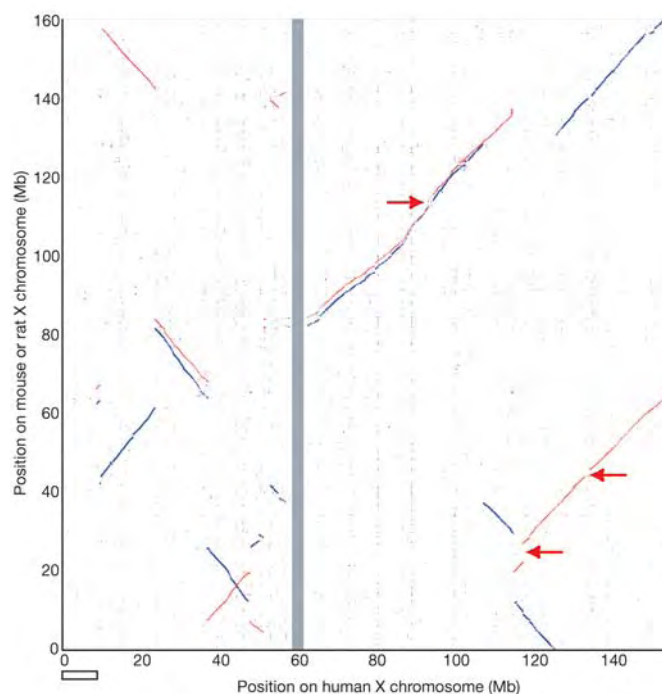


Figure 4 Conservation of the X chromosome in eutherian mammals. Plot of BLASTZ sequence alignments between the human X chromosome and the mouse (red) and rat (blue) X chromosomes. The rodent chromosomes are oriented with their centromeres pointing downwards. Regions indicated with arrows are long, highly similar repeats in the mouse sequence that are absent from the human and rat sequences. These repeats were apparently collapsed in an earlier analysed version of the mouse sequence, which also had a large inversion with respect to the mouse assembly used here (NCBI32)⁶⁶. The NCBI32 assembly has a gap from 0–3 Mb, which explains the absence of homology to the human X sequence in this part of the plot. The open horizontal bar shows the terminal section of human Xp, which is not conserved on the rodent X chromosomes.

rodent X chromosomes is the existence of 9 Mb of sequence at the tip of the human short arm (including human PAR1) that is apparently missing from the rodent X chromosomes (Fig. 4). There are 34 known and novel protein-coding genes in this segment of the human X chromosome (Supplementary Fig. 1), enabling us to investigate how this difference arose. A comprehensive database search of the rodent genome sequences revealed convincing orthologues for only thirteen of these genes in rat and five in mouse. Most of the rat orthologues are located in two groups on chromosome 12, and the only genes for which X-linked orthologues could be found in both rodents were *PRKX* and *STS*. In contrast, we found 24 of these 34 genes on chicken chromosome 1, and the order of these genes is perfectly conserved between the two genomes. Therefore, we conclude that this large terminal segment was present in the XAR and was subsequently removed from the X chromosome in a common murid ancestor of mouse and rat. The relative paucity of rodent ECRs in this segment of the X chromosome sequence (Supplementary Fig. 1) suggests that much of the region may be absent altogether from the genomes of *Mus musculus* and *Rattus norvegicus*.

Comparison of the human X and Y chromosomes

The evolutionary process has eradicated most traces of the ancestral relationship between the human X and Y chromosomes. At the cytogenetic level, the Y chromosome has a large and variably sized heterochromatic block and is considerably smaller than the

X chromosome, and the euchromatic part of the X chromosome is six times longer than that of Y. Few genes on human chromosome X have an active counterpart on the Y chromosome, and the majority of these are contained in regions where XY homology is of relatively recent origin.

A detailed comparison of the human X and Y chromosome sequences reveals the extent of Y chromosome decay in non-recombining regions. All of the large homologous blocks visible in Fig. 5 (and represented schematically in Fig. 6) are descended from material that was added to the established sex chromosomes. The tip of the short arm of the X and Y chromosomes comprises the 2.7 Mb pseudoautosomal region PAR1. Homology between the X and Y chromosomes in PAR1 is maintained by an obligatory recombination in male meiosis; gene loci in this region are present in two copies in both males and females and are not subject to dosage compensation by XCI. At the tip of the long arm of X and Y is a second pseudoautosomal region, the 330 kb PAR2, which was created by duplication of material from X to Y since the divergence of human and chimpanzee lineages⁴⁰. Some genes in PAR2 are subject to XCI, presumably reflecting their status on the X chromosome before the duplication event. Outside the PARs, homologies between the X and Y chromosomes are in non-recombining regions, predominantly in other parts of the XAR, together with a large 'X-transposed region' (XTR)⁴¹ in Xq21.3 and Yp11.2–p11.3 (see below). It is thought that the XAR originally formed a large pseudoautosomal region with an equivalent YAR, which is now

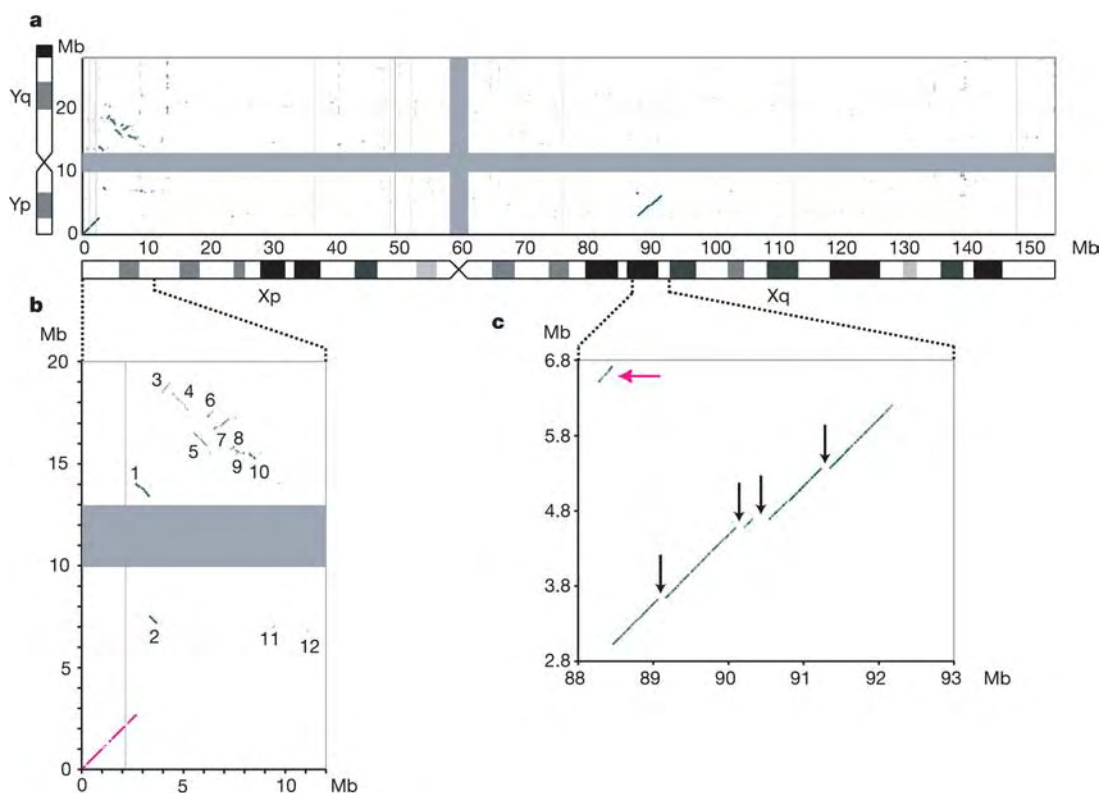


Figure 5 Limited homology between the human sex chromosomes illustrates the extent of Y chromosome erosion in non-recombining regions. **a**, BLASTN alignments (length ≥ 80 bp, sequence identity $\geq 70\%$) between the finished sequences of the X and Y chromosomes. The centromere positions are represented by grey bars. The analysed Y chromosome sequence ends at the large, heterochromatic segment on Yq, which is indicated by the black bar on the truncated Y chromosome ideogram. **b**, Major blocks of homology remaining between the XAR and the YAR. Expansion of the BLASTN plot from 0–12 Mb on the X chromosome and 0–20 Mb on the Y chromosome. On the X chromosome, the major homologies lie in the terminal 8.5 Mb of Xp: PAR1 (magenta

line) and numbered blocks 1–10. Lesser homologies 11 and 12 contain the *TBL1X/TBL1Y* and *AMELX/AMELY* genes, respectively. **c**, The XTR region in detail (88–93 Mb on X and 2.8–6.8 Mb on Y). Black arrows show large segments deleted from the Y chromosome copy of the XTR. The magenta arrow indicates the short segment that is separated from the rest of the XTR by a paracentric inversion on the Y chromosome. An independent inversion polymorphism on Yp in human populations encompasses this small segment. The position and orientation of the segment shows that the Y chromosome reference sequence is of the less common, derived Y chromosome.

largely eroded. At a gross level, the homology between the XAR and YAR is continuous for 6 Mb proximal to the pseudoautosomal boundary on the X (PABX), but is considerably more fragmented on the Y chromosome (Figs 5b and 6). Beyond this, the remaining 38.5 Mb of the XAR detects few other remnants of the YAR. Homologies are mostly in small islands around genes with functional orthologues on both sex chromosomes (for example, *AMELX/AMELY*, *ZFX/ZFY*, see Table 1).

The XTR arose by duplication of material from X to Y since the divergence of the human and chimpanzee lineages⁴². The duplicated region spans 3.91 Mb on X, but the corresponding region is only 3.38 Mb on the Y chromosome (Fig. 5c). We have aligned the entire X and Y copies of this region. Excluding insertions and deletions, sequence identity between the copies is 98.78%. We estimate that the transposition event occurred approximately 4.7 Myr ago (Supplementary Discussion 1), which is close to the suggested date of the speciation event that led to humans and chimpanzees, assumed here to be 6 Myr ago. The sequence alignment demonstrates the substantial changes to the XTR on the Y chromosome since the transposition. An inversion is known to have separated a 200-kb section from the rest of the XTR⁴³ (Fig. 5c). Also, the main block of homology is 540 kb shorter on Y than X, owing in particular to the absence of four large regions from the Y chromosome (Fig. 5c). The detection of these sequences at the expected positions on the chimpanzee X chromosome confirms that they were deleted from the Y chromosome after the transposition.

We found that only 54 of the 1,098 genes annotated on the X chromosome have functional homologues on the Y chromosome (Table 1). We obtained direct evidence for 24 genes in PAR1. Twenty-three of them are annotated (Supplementary Fig. 1), and the location of the 5' end of *CRLF2* indicates that the rest of this gene is in gap 4 of the human X sequence (see the VEGA database). On the basis of the excellent conservation of synteny between human PAR1 and the chicken sequence, we infer that a stromal antigen gene (orthologue of chicken Ensembl gene ENSGALG0000016716) lies in gap 1 (see Fig. 3b). As the annotated putative transcript cM56G10.2 might represent the 3' end of this gene, we conclude that PAR1 contains at least 24 genes. Together with the five annotated genes in PAR2, 29 genes lie entirely within the recombining regions of the sex chromosomes. Additionally, the XG locus spans the boundary between PAR1 and X-specific DNA, but has been disrupted by rearrangement on the Y chromosome.

Outside the XY-recombining regions of the X chromosome, we observed 25 genes that have functional homologues on the Y chromosome (Table 1). Fifteen of these are within the XAR, and a further three genes are shared by the X and Y copies of the XTR. The seven other XY gene pairs are believed to have descended from the proto-sex chromosomes. Only five cases have been described previously^{44,45}: the X chromosome genes are *SOX3*, *SMCX*, *RPS4X*, *RBMX* and *TSPYL2*, which are located on the long arm and proximal short arm (Table 1). The two additional cases we report here involve heat-shock transcription factor genes, designated *HSEFX1* and *HSEFX2*. They are assigned to the category of XCR genes on the basis of a high degree of divergence from their Y chromosome homologues and their location distal to *SOX3* within the XCR. *HSEFX1* and *HSEFX2* lie within the separate copies of a palindromic repeat in Xq28 and are identical to each other. By analogy, their Y chromosome homologues (*HSEFY1* and *HSEFY2*) lie within the arms of a Y chromosome palindrome, the similarity of which is thought to be maintained by gene conversion⁴¹.

On the basis of this and previously published information⁴¹, we can conclude that approximately 15 protein-coding genes on the Y chromosome have no detectable X chromosome homologue.

The progressive loss of XY recombination

The barrier to recombination between the proto-X and Y chromosomes initially encompassed the sex-determining locus on the

Y (*SRY*) and possibly other loci affecting male fitness. It is proposed that rearrangement of the Y chromosome led to the development of this barrier. Thereafter, successive rearrangements that encompassed parts of the pseudoautosomal region resulted in segments of Y-linked DNA that could no longer recombine and consequently degenerated over time. Evidence for the role of Y-specific (as opposed to X-specific) rearrangement in this phenomenon is most clearly illustrated by our analysis of the XAR, which shows very little rearrangement between human and avian lineages (Fig. 3a).

In a previous study⁴⁶, four broad physical and evolutionary regions were defined on the X chromosome. The X chromosome genes within a given region all showed a similar level of divergence from their Y chromosome counterparts. However, between regions, levels of divergence were very different, presumably reflecting the stepwise loss of recombination between the X and Y chromosomes.

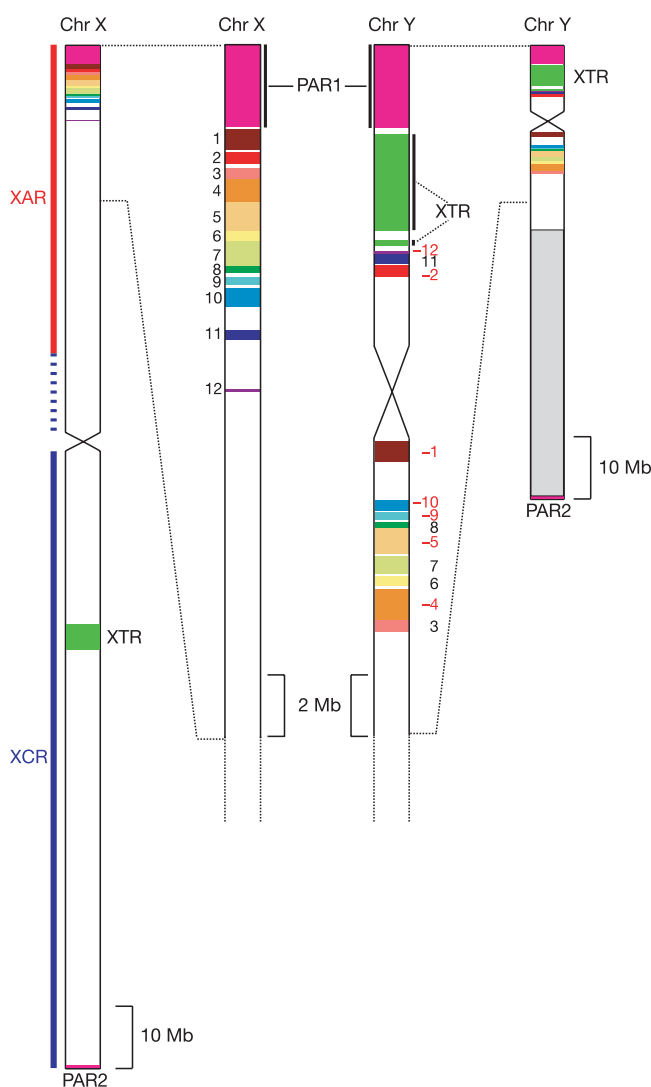


Figure 6 Schematic representation of major homologies between the human sex chromosomes. The entire X and Y chromosomes are shown using the same scale on the left and right sides of the figure, respectively. The major heterochromatic region on Yq is indicated by the pale grey box proximal to PAR2. Expanded sections of X and Y are shown in the centre of the figure. Homologies coloured in the figure are either part of the XAR (PAR1 and blocks 1–12), or were duplicated from the X chromosome to the Y chromosome since the divergence of human and chimpanzee lineages (XTR and PAR2). The numbering of XAR–YAR blocks follows that in Fig. 5b. Blocks inverted on the Y chromosome relative to the X chromosome are assigned red, negative numbers.

The physical order of the four regions on the X chromosome was seen to parallel their evolutionary ages, and therefore the chromosome was described as having four “evolutionary strata”⁴⁶. In general, gene pairs were found to be less divergent moving through the strata from Xqter to Xpter. The first two strata (S1 and S2) encompass the long arm and proximal short arm, respectively, and were defined by the genes that survive from the proto-sex chromosomes. Gene pairs were found to be increasingly similar moving through strata 3 and 4, which occupy the proximal and distal sections of the XAR, respectively.

We re-evaluated XY homology in S4 and S3 using finished, genomic sequences from the two chromosomes. For S4 in particular, substantial blocks of homology exist between the chromosomes (blocks 1–10 in Fig. 5b and Fig. 6). Aligning the X and Y chromosome sequences across this region, we observed a bipartite organization, with markedly greater XY identity in the distal 1.0 Mb compared with the proximal 4.5 Mb (Fig. 7a). On this basis, the

distal portion containing the *GYG2*, *ARSD*, *ARSE*, *ARSE*, *ADLICAN* and *PRKX* genes can be redefined as a new, fifth stratum, S5 (Figs 1 and 7a). A most parsimonious series of inversions, from the current arrangement of homologous blocks on X to that on Y, is consistent with the proposed strata (Fig. 7b). These data refine the picture of loss of XY recombination during evolution, which occurred by migration of the PABX in a stepwise manner distally through the XAR. The available evidence now suggests that there have been at least four PABX positions within the XAR, which are at the S2/S3, S3/S4 and S4/S5 boundaries (~47 Mb, ~8.5 Mb and ~4 Mb from Xpter, respectively), and at the current position (2.7 Mb from Xpter). We estimate that the two most recent PABX movements, which created first S4 and then S5, occurred 38–44 Myr ago and 29–32 Myr ago, respectively (Supplementary Discussion 2).

In addition to the varied degree of XY sequence identity within S3, S4, S5 and PAR1, we found marked differences in their sequence composition, which were presumably also caused by the loss of

Table 1 Homologous genes on the human X and Y chromosomes

Region	Distance from Xpter (Mb)	X gene*	Y gene	Distance from Ypter (Mb)†	XY homology block‡
Pseudoautosomal region PAR1 (XAR)	0.15	cU136G2.1 (<i>PLCXD1</i>)	cU136G2.1 (<i>PLCXD1</i>)	0.15	PAR1
	0.17	cU136G2.2 (<i>GTPBP6</i>)	cU136G2.2 (<i>GTPBP6</i>)	0.17	PAR1
	0.25	cM56G10.2§	cM56G10.2§	0.25	PAR1
	0.29	cM56G10.1 (<i>PPP2R3B</i>)	cM56G10.1 (<i>PPP2R3B</i>)	0.29	PAR1
	0.57	<i>SHOX</i>	<i>SHOX</i>	0.57	PAR1
	0.92	bA309M23.1§	bA309M23.1§	0.92	PAR1
	1.31	<i>CRLF2</i>	<i>CRLF2</i>	1.31	PAR1
	1.38	<i>CSF2RA</i>	<i>CSF2RA</i>	1.38	PAR1
	1.52	<i>IL3RA</i>	<i>IL3RA</i>	1.52	PAR1
	1.55	<i>SLC25A6</i>	<i>SLC25A6</i>	1.55	PAR1
	1.56	bA261P4.5§	bA261P4.5§	1.56	PAR1
	1.57	bA261P4.6 (<i>CXYorf2</i>)	bA261P4.6 (<i>CXYorf2</i>)	1.57	PAR1
	1.59	<i>ASMTL</i>	<i>ASMTL</i>	1.59	PAR1
	1.66	bA261P4.4 (<i>P2RY8</i>)	bA261P4.4 (<i>P2RY8</i>)	1.66	PAR1
	1.76	<i>DXYS155E</i> (<i>CXYorf3</i>)	<i>DXYS155E</i> (<i>CXYorf3</i>)	1.76	PAR1
	1.79	<i>ASMT</i>	<i>ASMT</i>	1.79	PAR1
	1.79	bB297E16.3§	bB297E16.3§	1.79	PAR1
	1.91	bB297E16.4§	bB297E16.4§	1.91	PAR1
	1.93	bB297E16.5§	bB297E16.5§	1.93	PAR1
	2.37	<i>DHRSX</i>	<i>DHRSX</i>	2.37	PAR1
	2.41	<i>ALTE</i> (<i>ZBED1</i>)	<i>ALTE</i> (<i>ZBED1</i>)	2.41	PAR1
	2.54	Em:AC097314.2§	Em:AC097314.2§	2.54	PAR1
	2.53	Em:AC097314.3§	Em:AC097314.3§	2.53	PAR1
	2.63	<i>CD99</i>	<i>CD99</i>	2.63	PAR1
	3.57	<i>PRKX</i>	<i>PRKY</i>	7.23	2
	5.81	<i>NLGN4X</i>	<i>NLGN4Y</i>	15.23	5
	6.31	Em:AC108684.1 (<i>VCX3A</i>)	<i>VCY</i> , <i>VCY1B</i>	14.54, 14.6	6
	7.62	<i>VCX</i>	<i>VCY</i> , <i>VCY1B</i>	14.54, 14.6	9
	7.95	Em:AC097626.1 (<i>VCX2</i>)	<i>VCY</i> , <i>VCY1B</i>	14.54, 14.6	10
	8.24	Em:AC006062.2 (<i>VCX3B</i>)	<i>VCY</i> , <i>VCY1B</i>	14.54, 14.6	10
	9.37	<i>TBL1X</i>	<i>TBL1Y</i>	6.97	11
	11.07	<i>AMELX</i>	<i>AMELY</i>	6.78	12
	12.75	<i>TMSB4X</i>	<i>TMSB4Y</i>	14.25	
	16.59	<i>CXorf15</i>	<i>CYorf15A</i> , <i>CYorf15B</i>	20.13, 20.15	
	19.91	<i>EIF1AX</i>	<i>EIF1AY</i>	21.08	
	23.96	<i>ZFX</i>	<i>ZFY</i>	2.87	
	40.78	<i>USP9X</i>	<i>USP9Y</i>	13.33	
	40.96	<i>DDX3X</i>	<i>DDX3Y</i>	13.46	
	44.61	<i>UTX</i>	<i>UTY</i>	13.91	
X-conserved region (XCR)	53.00	dJ290F12.2 (<i>TSPYL2</i>)	<i>TSPY</i> (~35)	9.50	
	53.12	<i>SMCX</i>	<i>SMCY</i>	20.27	
	71.27	<i>RPS4X</i>	<i>RPS4Y1</i> , <i>RPS4Y2</i>	2.77, 21.27	
X-transposed region (XTR)	88.50	bB348B13.2§	n/a	2.96	XTR
	88.99	<i>TGIF2LX</i>	<i>TGIF2LY</i>	3.49	XTR
	91.26	<i>PCDH11X</i>	<i>PCDH11Y</i>	5.28	XTR
X-conserved region (XCR)	135.68	<i>RNIMX</i> (<i>RBMX</i>)	<i>RBMX</i> (6)	22.02, 22.04, 22.37, 22.41, 22.66, 22.85	
	139.31	<i>SOX3</i>	<i>SRY</i>	2.70	
	148.38	Em:AC016940.3 (<i>HSFX2</i>)§	<i>HSFY1</i> , <i>HSFY2</i>	19.3, 19.12	
	148.56	Em:AC016939.4 (<i>HSFX1</i>)§	<i>HSFY1</i> , <i>HSFY2</i>	19.3, 19.12	
Pseudoautosomal region PAR2	154.57	<i>SPRY3</i>	<i>SPRY3</i>	57.44	PAR2
	154.71	<i>SYBL1</i>	<i>SYBL1</i>	57.58	PAR2
	154.81	<i>IL9R</i>	<i>IL9R</i>	57.67	PAR2
	154.81	Em:AJ271736.5§	Em:AJ271736.5§	57.69	PAR2
	154.82	Em:AJ271736.6 (<i>FAM39A</i>)§	Em:AJ271736.6 (<i>FAM39A</i>)§	57.69	PAR2

Pseudogenes are not included in the table.

*Gene names as shown in Supplementary Fig. 1. HUGO name is in parentheses when the two names differ. Em, EMBL entry.

†Distances refer to Y chromosome sequence assembly NCBI35. Where multiple Y chromosome orthologues exist, the locations of all copies are shown on the Y chromosome. The exception is *TSPY*, which has ~35 copies in an array centred at approximately 9.5 Mb on the Y chromosome⁴¹.

‡Major homology blocks as shown in Figs 5 and 6.

§Novel cases of X genes with Y homologues assigned to these categories.

recombination in each region during evolution. Specifically, we observed that L1, L2 and mammalian interspersed repeat (MIR) coverage decrease with each more distal stratum and PAR1 (Table 2 and Fig. 1), but (G+C) levels and *Alu* repeat content increase abruptly at the boundary between S4 and S5 (Table 2 and Fig. 8); variations in the incidence of different *Alu* subfamilies (Y, S and J) also contribute to the distinct character of each stratum and PAR1 (Supplementary Table 13). The compositional differences between S4 and S5 provide additional support for the subdivision of the original stratum 4 (Fig. 8).

X-chromosome inactivation

XCI in mammals achieves dosage compensation between males and females for X-linked gene products. Inactivation of one X chromosome occurs early in female development and is initiated from the X-inactivation centre (XIC). The *XIST* transcript is expressed initially on both X chromosomes, but later the transcript from the chromosome that is destined for inactivation becomes more stable than the other. Finally, the transcript is expressed only from the inactive X chromosome (X_i). Coating with the *XIST*

transcript is the earliest of many chromatin modifications on X_i .

XCI was first proposed based partly on the study of X:autosome translocations in female mice⁴⁷. Studies of derivative chromosomes containing inactivated X chromosome segments later concluded that the inactivation could spread across the translocation boundary to the autosomal segment, but that inactivation of this segment was incomplete. More recently, it has become clear that more than 15% of the genes on the human X chromosome, including many without functional equivalents on the Y, escape from XCI, as presented in detail elsewhere⁴⁸. The majority of the genes that escape XCI lie within the distal regions of the XAR (Fig. 1): all genes studied in PAR1, S5 and S4 were found to escape from XCI, but there is a lower proportion of escapees in S3, and very few examples in the XCR⁴⁸. This observation correlates with our picture of X chromosome evolution: XCI follows Y chromosome attrition⁴⁹, which is less advanced in the distal strata of the XAR.

Inefficient inactivation of the autosomal segment in X_i :autosome translocations led to the proposal that ‘way stations’ on the X chromosome boost the spread of XCI. According to this model, way stations are present throughout the genome but are enriched on the

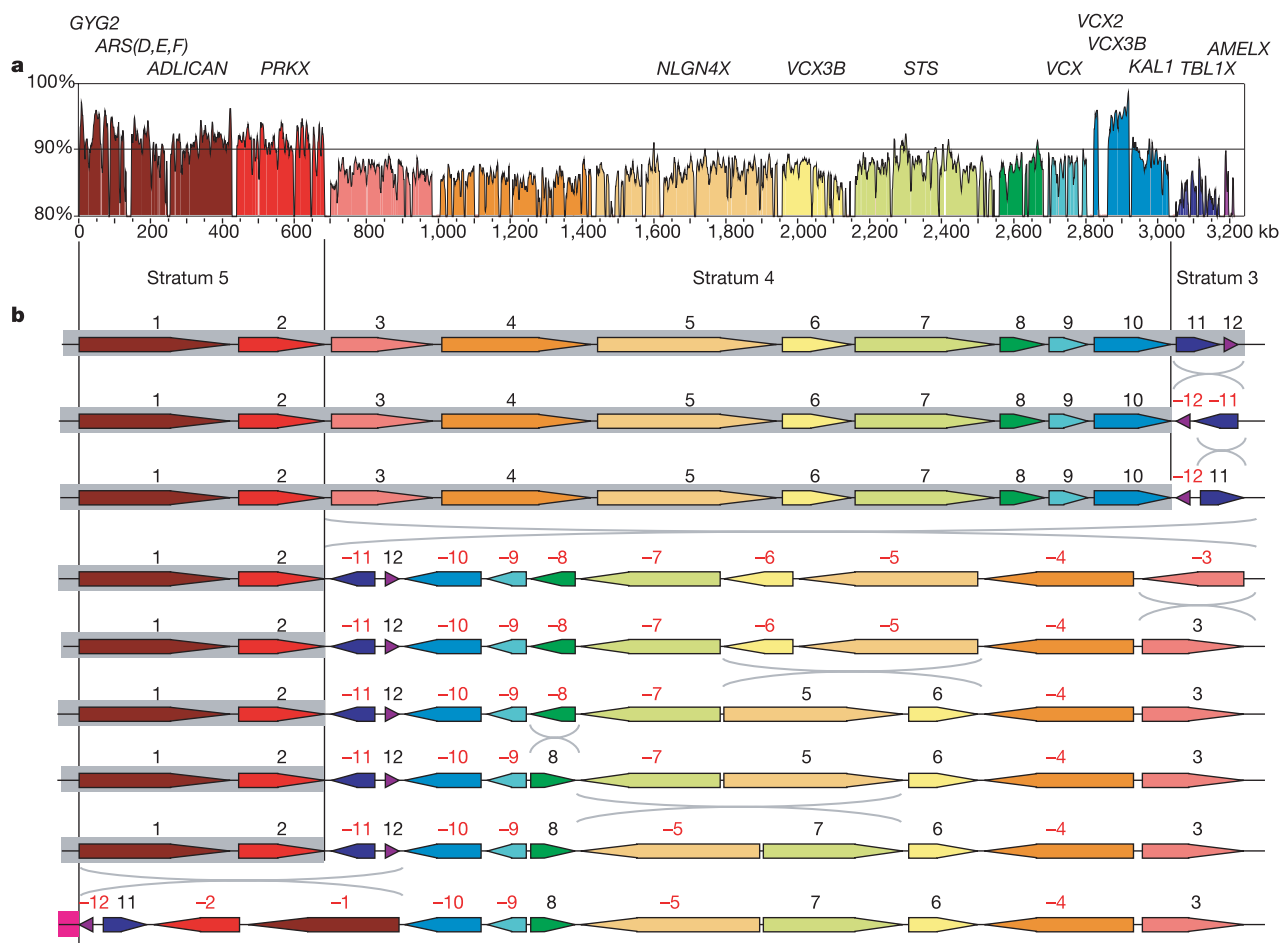


Figure 7 Evidence for a fifth evolutionary stratum on the X chromosome. **a**, Sequence identity between the X and Y homology blocks 1–12 (see Figs 5b and 6) plotted in 5-kb windows. The scale shows the total amount of sequence aligned, excluding insertions and deletions (see Methods). A 10-kb spacer is placed between each consecutive block of homology. Segments of the plot are coloured according to the system used in Figs 6 and 7b. On the basis of this plot, a new evolutionary stratum S5 is defined, which includes homology blocks 1 and 2. **b**, A most parsimonious series of inversion events from the arrangement of homology blocks 1–12 on the X chromosome (top) to the Y chromosome (bottom), calculated using GRIMM⁶⁴. The grey boxes show the suggested extents of former pseudoautosomal regions within the distal part of the XAR, and the magenta box

(bottom row) shows the position of the current pseudoautosomal region. This inversion sequence provides independent support for the proposed pseudoautosomal boundary movements and evolutionary strata. It was previously suggested that *AMELX* (in block 12) is in S4 (ref. 46), or possibly at the boundary between S3 and S4 (ref. 67). However, the more distal location of block 11, which contains *TBL1X* (an S3 gene⁴⁶), is not consistent with these suggestions. The two regions of increased sequence identity within block 10 contain the *VCX2* and *VCX3B* genes on the X chromosome and the *VCY1B* and *VCY* genes on the Y chromosome. This gene family might have arisen *de novo* in the simian lineage⁶⁸, which could account for the unusual characteristics of this part of the alignment.

Table 2 Sequence characteristics of evolutionary domains of the X chromosome

Region	(G+C) (%)	L1 (%)	L1P (%)	L1M (%)	Alu (%)	L2 (%)	MIR+MIR3 (%)
X chromosome	39.46	28.87	13.39	15.21	8.23	2.98	2.07
XAR	39.87	17.89	6.60	11.23	10.28	2.63	1.76
XCR	39.28	33.50	16.38	16.97	7.28	3.12	2.19
PAR1	48.11	6.97	2.64	4.38	28.88	0.24	0.21
S5	42.86	8.89	4.36	4.59	18.72	0.66	0.31
S4	38.87	11.10	4.31	6.80	8.60	1.59	0.95
S3	39.46	19.55	7.14	12.34	9.24	2.94	1.98

See Supplementary Table 13 for additional repeat element data.

X chromosome, particularly in the region of the XIC⁵⁰. Lyon suggested that L1 elements are good candidates for acting as way stations on account of their enrichment on the mammalian X chromosome⁵¹. We observe a distribution of L1 elements on the chromosome that is consistent with both the way station and the Lyon hypotheses (Fig. 1 and Table 2). The coverage of L1 repeats is very high in the XCR, especially around the XIC. As noted previously⁵², this enrichment in L1 levels is accounted for particularly by elements that were active more recently in mammalian evolution⁵³ (L1P in Fig. 1). In the XAR, L1 coverage is close to autosomal levels, whereas L1 levels are particularly low in the distal evolutionary strata of the XAR, where genes consistently escape inactivation. The *XIST* locus itself lies in a 60 kb region that is virtually devoid of L1 elements, whereas L1 levels are extremely high in the adjacent regions. Based on their distributions, other interspersed repeats are not strong candidates for way stations. For example, although L2 and MIR elements are reduced in S4, S5 and particularly PAR1 relative to the rest of the chromosome, their overall levels on the X chromosome are not enriched relative to the autosomes but are slightly reduced. Furthermore, L2 and MIR levels are low in the region distal to the XIC. These characteristics do not preclude an involvement in XCI, but are not consistent with a role as way stations.

The possible causal relationship of L1 elements to the spread of XCI remains a subject of debate. Some studies have reported significant associations between L1 coverage and inactivation⁵², and others have refuted this⁵⁴. Our observations on regional differences in composition emphasize that such studies should compare active and inactivated genes (or domains) from the same evolutionary stratum, in order to avoid correlations that are unrelated to XCI.

Medical genetics and the X chromosome sequence

The X chromosome holds a unique place in the history of medical genetics. Ascertainment of X-linked diseases is enhanced by the relative ease of recognizing this mode of inheritance. More

important, however, is the fact that a disproportionately large number of disease conditions have been associated with the X chromosome because the phenotypic consequence of a recessive mutation is revealed directly in males for any gene that has no active counterpart on the Y chromosome. Thus, although the X chromosome contains only 4% of all human genes, almost 10% of diseases with a mendelian pattern of inheritance have been assigned to the X chromosome (307 out of 3,199; information obtained from OMIM²). These two aspects of the medical genetics of the X chromosome have greatly stimulated progress in the positional cloning of many genes associated with human disease. To date, the molecular basis for 168 X-linked phenotypes has been determined, and the X chromosome sequence has aided this process for 43 of them, by providing positional candidate genes or a reference sequence for comparison to patient samples (Supplementary Table 14).

Identifying genes involved in rare conditions yields important biological insights. For example, discovery of mutations in the *SH2D1A* gene⁵⁵ (involved in X-linked lymphoproliferative disease (XLP, OMIM 308240)) led to identification of a new mediator of signal transduction between T and NK cells, and a novel family of proteins involved in the regulation of the immune response. Mental retardation is one of the most common problems in clinical genetics, and affects significantly more males than females. To date, 16 genes from the X chromosome have been associated with cases of non-syndromic X-linked mental retardation (NS-XLMR), in which mental retardation is the only phenotypic feature. These genes encode a range of protein types, and some are also involved in syndromic forms of mental retardation. For example, the *ARX* gene encodes an aristaless-related homeobox transcription factor and is linked to NS-XLMR cases, as well as to syndromic mental retardation associated with epilepsy (infantile spasm syndrome, ISSX, OMIM 308350) or with dystonic hand movements (Partington syndrome, PRTS, OMIM 309510)⁵⁶. The *MECP2* gene, which encodes a methyl-CpG-binding protein, was initially linked to cases of Rett syndrome in girls⁵⁷ (RTT, OMIM 312750) but was later also seen to be mutated in males or females with NS-XLMR⁵⁸. The molecular defect has been determined in only a minority of families affected by NS-XLMR, which has led to speculation that there could be as many as 100 genes on the X chromosome that are associated with NS-XLMR⁵⁹. Discovering the genes for these and other rare, monogenic disorders is of critical value in extending our understanding of fundamental new processes in human biology, and the annotated X chromosome will further facilitate this process.

Concluding remarks

The completion of the X chromosome sequencing project is an essential component of the goal of obtaining a high-quality,

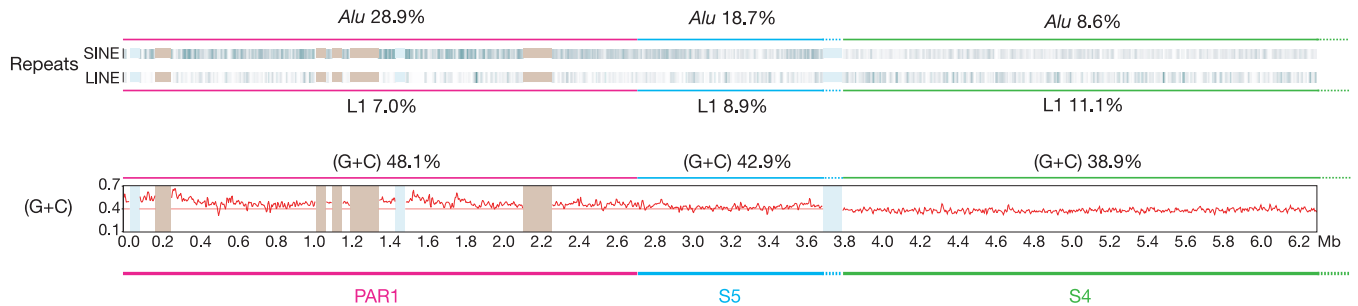


Figure 8 Sequence compositional changes in the distal evolutionary strata of the X chromosome. Shown are the positions of SINE and LINE repeats and (G+C) content within PAR1, S5 and the distal half of S4. The percentage of *Alu*, L1 and (G+C) are shown for each region (including the whole of S4). There is an abrupt increase in *Alu* repeat levels

and (G+C) content from S4 to S5. The five euchromatic gaps in PAR1 are shown as light brown bars. Pale blue bars represent clones for which the sequences were unfinished at the time of the sequence assembly.

annotated human genome sequence for use in studies of gene function, sequence variation, disease and evolution. It also means that for the first time, we now have the finished sex chromosome sequences of an organism. The study of these sequences gives a greater insight into mammalian sex chromosome evolution and its consequences. As these analyses are extended to other genomes, we will gain a greater appreciation of the different evolutionary forces that shape sex chromosome and autosome evolution. It will be important to study differences in the rates of mutational processes, and to consider the influence of the unusual pattern of male recombination on these processes. Clearly, this analysis should not be restricted to a consideration of mammalian sex chromosomes, and it will be of great interest to make comparisons with non-mammalian systems that arose independently in evolution. □

Methods

The approach used to establish a bacterial clone map of the X chromosome has been previously described⁵. 13,264 clones were identified using 4,363 STS markers derived from published genetic or physical maps, from shotgun sequencing of flow-sorted X chromosomes, or from end-sequences of clones at contig ends. Clones were assembled into contigs using restriction-enzyme fingerprinting, and were integrated with the Washington University Genome Sequencing Center whole genome BAC map⁶⁰ in order to identify additional clones. Nine euchromatic gaps were measured using fluorescent *in situ* hybridization of clones to extended DNA fibres, and a tenth gap was estimated on the basis of end-sequence data from spanning, unstable BAC clones (Supplementary Table 2). On the basis of pulsed-field gel electrophoresis experiments, we expect the sizes of the other four euchromatic gaps to have a combined size of less than 400 kb.

Finished sequences of individual clones were determined using procedures described in ref. 7. For the analyses described above, the sequence was frozen in March 2004, at which point 150,396,262 bp of sequence had been determined from a minimal tiling path of 1,832 clones (1,616 sequence accessions). This sequence is available at <http://www.sanger.ac.uk/HGP/ChrX/>, and its annotation is represented in Supplementary Fig. 1. Updates to the sequence and annotation can be obtained from the VEGA database.

Manual annotation of gene structures has been described elsewhere¹⁴, and used guidelines agreed at the human annotation workshop (HAWK; <http://www.sanger.ac.uk/HGP/havana/hawk.shtml>). Genes were assigned to one of four groups: (1) known genes that are identical to human cDNAs or protein sequences and have a RefSeq RNA (and RefSeq protein, if the gene encodes a protein); (2) novel coding sequences, which have an open reading frame (ORF) and are identical to spliced ESTs, or have similarity to other genes/proteins (any species); (3) novel transcripts, which are similar to novel coding sequences, except that no ORF can be determined with confidence; and (4) putative transcripts, which are identical to splicing human ESTs but have no ORF. Gene symbols were approved by the HUGO Gene Nomenclature Committee wherever possible. Predicted protein translations were analysed for Pfam domains using InterProScan (<http://www.ebi.ac.uk/InterProScan/>). CpG islands were predicted using the program GpG (G. Micklem, personal communication).

Interspersed repeats were identified and classified using RepeatMasker (<http://repeatmasker.genome.washington.edu>). In order to search for segmental duplications, WU-BLASTN (<http://blast.wustl.edu>) was used to align the current X chromosome sequence to itself or to the NCBI34 autosome assemblies. Duplicated blocks at least 5 kb in length were defined as described in ref. 28.

SNPs (dbSNP release 119) were mapped onto the X chromosome sequence using first SSAHA⁶¹ and then Cross-match (<http://www.phrap.org/phredphrapconsd.html>).

Comparative analysis

The genome assemblies used for comparative analyses were: *Gallus gallus* WASHUC1 (Washington University Genome Sequencing Center, <http://www.genome.wustl.edu/projects/chicken/>), *Rattus norvegicus* RGSC3.1 (Rat Genome Sequencing Consortium <http://www.hgsc.bcm.tmc.edu/projects/rat/>), *Mus musculus* NCBI32 (Mouse Genome Sequencing Consortium, <http://www.ncbi.nlm.nih.gov/genome/seq/NCBIContigInfo.html>), *Danio rerio* version 3 (Sanger Institute, http://www.sanger.ac.uk/Projects/D_rerio/), *T. nigroviridis* version 6 (Genoscope and the Broad Institute, <http://www.genoscope.cns.fr/externe/tetraodon/Ressource.html>), and *F. rubripes* version 2 (International Fugu Genome Consortium, <http://www.fugu-sg.org/project/info.html>). ECRs between the X chromosome and the rodent and fish genomes were obtained as described elsewhere¹³. In order to visualize regions of conserved synteny, the X chromosome sequence was aligned to the chicken and rodent genome sequences using BLASTZ (with default parameters), and matches were plotted by chromosome position. Matches to the rodent genomes were filtered to include only those with a sequence identity of at least 70% to the human sequence. The Ensembl database (<http://www.ensembl.org>) was used to search for orthologous gene pairs between the X chromosome and the other three genomes.

Genomic sequence homologies between the X and Y chromosomes were identified by aligning the two finished chromosome sequences using WU-BLASTN, and then filtering the alignments to include only those of at least 70% sequence identity and 80 bp length. In order to calculate the sequence identity between large, XY-homologous regions, a global alignment of unmasked sequence was generated using LAGAN⁶². Gapped regions, which result from insertions or deletions, were removed from the alignment, and then the nucleotide sequence identity was calculated for the remainder. Sequence identity plots

were produced by parsing the LAGAN output into VISTA⁶³. GRIMM⁶⁴ was used to calculate a most parsimonious series of inversions that would account for differences in homology block order and orientation between the X and Y chromosomes. Homologous protein-coding gene pairs between the X and Y chromosomes were identified by TBLASTN searching with the coding sequences of annotated coding genes on the Y chromosome against the X chromosome genomic sequence.

Received 1 February; accepted 7 February 2005; doi:10.1038/nature03440.

- Ohno, S. *Sex Chromosomes and Sex-linked Genes* (Springer, Berlin, 1967).
- McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University and National Center for Biotechnology Information, National Library of Medicine. OMIM: *Online Mendelian Inheritance in Man*. (<http://www.ncbi.nlm.nih.gov/omim/>) (2000).
- Royer-Pokora, B. *et al.* Cloning the gene for an inherited human disorder—chronic granulomatous disease—on the basis of its chromosomal location. *Nature* **322**, 32–38 (1986).
- Monaco, A. P. *et al.* Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene. *Nature* **323**, 646–650 (1986).
- Bentley, D. R. *et al.* The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X. *Nature* **409**, 942–943 (2001).
- Humphray, S. J. *et al.* DNA sequence and analysis of human chromosome 9. *Nature* **429**, 369–374 (2004).
- International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
- Schmutz, J. *et al.* Quality assessment of the human genome sequence. *Nature* **429**, 365–368 (2004).
- Mahtani, M. M. & Willard, H. F. Physical and genetic mapping of the human X chromosome centromere: repression of recombination. *Genome Res.* **8**, 100–110 (1998).
- Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nature Genet.* **31**, 241–247 (2002).
- Pruitt, K. D., Katz, K. S., Sicotte, H. & Maglott, D. R. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet.* **16**, 44–47 (2000).
- Mungall, A. J. *et al.* The DNA sequence and analysis of human chromosome 6. *Nature* **425**, 805–811 (2003).
- Deloukas, P. *et al.* The DNA sequence and comparative analysis of human chromosome 10. *Nature* **429**, 375–381 (2004).
- Dunham, A. *et al.* The DNA sequence and analysis of human chromosome 13. *Nature* **428**, 522–528 (2004).
- Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S. R. Rfam: an RNA family database. *Nucleic Acids Res.* **31**, 439–441 (2003).
- Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
- Griffiths-Jones, S. The microRNA Registry. *Nucleic Acids Res.* **32** (Database issue), D109–111 (2004).
- Brown, C. J. *et al.* A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature* **349**, 38–44 (1991).
- Brown, C. J. *et al.* The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* **71**, 527–542 (1992).
- Lee, J. T., Davidow, L. S. & Warshawsky, D. Tsix, a gene antisense to Xist at the X-inactivation centre. *Nature Genet.* **21**, 400–404 (1999).
- Migeon, B. R., Chowdhury, A. K., Dunston, J. A. & McIntosh, I. Identification of TSIX, encoding an RNA antisense to human XIST, reveals differences from its murine counterpart: implications for X inactivation. *Am. J. Hum. Genet.* **69**, 951–960 (2001).
- Chow, J. C., Hall, L. L., Clemson, C. M., Lawrence, J. B. & Brown, C. J. Characterization of expression at the human XIST locus in somatic, embryonal carcinoma, and transgenic cell lines. *Genomics* **82**, 309–322 (2003).
- Chureau, C. *et al.* Comparative sequence analysis of the X-inactivation center region in mouse, human, and bovine. *Genome Res.* **12**, 894–908 (2002).
- Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **32** (Database issue), D138–141 (2004).
- Scanlan, M. J., Simpson, A. J. & Old, L. J. The cancer/testis genes: review, standardization, and commentary. *Cancer Immun.* [online] **4**, 1 (2004).
- Hurst, L. D. Evolutionary genomics: Sex and the X. *Nature* **411**, 149–150 (2001).
- Chomez, P. *et al.* An overview of the MAGE gene family with the identification of all human members of the family. *Cancer Res.* **61**, 5544–5551 (2001).
- Cheung, J. *et al.* Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol.* **4**, R25 (2003).
- Stankiewicz, P. & Lupski, J. R. Genome architecture, rearrangements and genomic disorders. *Trends Genet.* **18**, 74–82 (2002).
- Pegoraro, E. *et al.* Familial skewed X inactivation: a molecular trait associated with high spontaneous abortion rate maps to Xq28. *Am. J. Hum. Genet.* **61**, 160–170 (1997).
- Schueler, M. G., Higgins, A. W., Rudd, M. K., Gustashaw, K. & Willard, H. F. Genomic and genetic definition of a functional human centromere. *Science* **294**, 109–115 (2001).
- Spence, J. M. *et al.* Co-localization of centromere activity, proteins and topoisomerase II within a subdomain of the major human X alpha-satellite array. *EMBO J.* **21**, 5269–5280 (2002).
- Rudd, M. K. & Willard, H. F. Analysis of the centromeric regions of the human genome assembly. *Trends Genet.* **20**, 529–533 (2004).
- She, X. *et al.* The structure and evolution of centromeric transition regions within the human genome. *Nature* **430**, 857–864 (2004).
- The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
- Schmid, M. *et al.* First report on chicken genes and chromosomes 2000. *Cytogenet. Cell Genet.* **90**, 169–218 (2000).
- Kohn, M., Kehr-Sawatzki, H., Vogel, W., Graves, J. A. & Hameister, H. Wide genome comparisons reveal the origins of the human X chromosome. *Trends Genet.* **20**, 598–603 (2004).
- Graves, J. A. The origin and function of the mammalian Y chromosome and Y-borne genes—an evolving understanding. *Bioessays* **17**, 311–320 (1995).
- Hillier, L. W. *et al.* Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695–716 (2004).

40. Freije, D., Helms, C., Watson, M. S. & Donis-Keller, H. Identification of a second pseudoautosomal region near the Xq and Yq telomeres. *Science* **258**, 1784–1787 (1992).
41. Skaletsky, H. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825–837 (2003).
42. Page, D. C., Harper, M. E., Love, J. & Botstein, D. Occurrence of a transposition from the X-chromosome long arm to the Y-chromosome short arm during human evolution. *Nature* **311**, 119–123 (1984).
43. Sargent, C. A. *et al.* The sequence organization of Yp/proximal Xq homologous regions of the human sex chromosomes is highly conserved. *Genomics* **32**, 200–209 (1996).
44. Toder, R., Wakefield, M. J. & Graves, J. A. The minimal mammalian Y chromosome - the marsupial Y as a model system. *Cytogenet. Cell Genet.* **91**, 285–292 (2000).
45. Delbridge, M. L. *et al.* TSPY, the candidate gonadoblastoma gene on the human Y chromosome, has a widely expressed homologue on the X - implications for Y chromosome evolution. *Chromosome Res.* **12**, 345–356 (2004).
46. Lahn, B. T. & Page, D. C. Four evolutionary strata on the human X chromosome. *Science* **286**, 964–967 (1999).
47. Lyon, M. F. Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature* **190**, 372–373 (1961).
48. Carrel, L. & Willard, H. F. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* doi:10.1038/nature03479 (this issue).
49. Jegalian, K. & Page, D. C. A proposed path by which genes common to mammalian X and Y chromosomes evolve to become X inactivated. *Nature* **394**, 776–780 (1998).
50. Gartler, S. M. & Riggs, A. D. Mammalian X-chromosome inactivation. *Annu. Rev. Genet.* **17**, 155–190 (1983).
51. Lyon, M. F. X-chromosome inactivation: a repeat hypothesis. *Cytogenet. Cell Genet.* **80**, 133–137 (1998).
52. Bailey, J. A., Carrel, L., Chakravarti, A. & Eichler, E. E. Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. *Proc. Natl Acad. Sci. USA* **97**, 6634–6639 (2000).
53. Smit, A. F., Toth, G., Riggs, A. D. & Jurka, J. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J. Mol. Biol.* **246**, 401–417 (1995).
54. Ke, X. & Collins, A. CpG islands in human X-inactivation. *Ann. Hum. Genet.* **67**, 242–249 (2003).
55. Coffey, A. J. *et al.* Host response to EBV infection in X-linked lymphoproliferative disease results from mutations in an SH2-domain encoding gene. *Nature Genet.* **20**, 129–135 (1998).
56. Stromme, P. *et al.* Mutations in the human ortholog of *Aristaless* cause X-linked mental retardation and epilepsy. *Nature Genet.* **30**, 441–445 (2002).
57. Amir, R. E. *et al.* Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nature Genet.* **23**, 185–188 (1999).
58. Orrico, A. *et al.* MECP2 mutation in male patients with non-specific X-linked mental retardation. *FEBS Lett.* **481**, 285–288 (2000).
59. Ropers, H. H. *et al.* Nonsyndromic X-linked mental retardation: where are the missing mutations? *Trends Genet.* **19**, 316–320 (2003).
60. The International Human Genome Mapping Consortium. A physical map of the human genome.

Nature **409**, 934–941 (2001).

61. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11**, 1725–1729 (2001).
62. Brudno, M. *et al.* LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13**, 721–731 (2003).
63. Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M. & Dubchak, I. VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* **32**, W273–W279 (2004).
64. Bourque, G. & Pevzner, P. A. Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Res.* **12**, 26–36 (2002).
65. Francke, U. Digitized and differentially shaded human chromosome ideograms for genomic applications. *Cytogenet. Cell Genet.* **65**, 206–218 (1994).
66. Gibbs, R. A. *et al.* Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521 (2004).
67. Iwase, M. *et al.* The amelogenin loci span an ancient pseudoautosomal boundary in diverse mammalian species. *Proc. Natl Acad. Sci. USA* **100**, 5258–5263 (2003).
68. Lahn, B. T. & Page, D. C. A human sex-chromosomal gene family expressed in male germ cells and encoding variably charged proteins. *Hum. Mol. Genet.* **9**, 311–319 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank the Washington University Genome Sequencing Center for access to chicken and chimpanzee genome sequence data before publication; the Broad Institute for access to dog and chimpanzee genome sequence data before publication, for *T. nigroviridis* genome data, and for fosmid end-sequence data and clones; members of the Sanger Institute zebrafish genome project; the mouse, rat and *F. rubripes* sequencing consortia; Genoscope for *T. nigroviridis* genome data; the Ensembl, UCSC, EMBL and GenBank database groups; G. Schuler for information on sequence overlaps; T. Furey for information on RefSeq RNA coverage; D. Jaffe for data on fosmid end-sequence matches; D. Vetrie, E. Kendall, D. Stephan, J. Trent, A. P. Monaco, J. Chelly, D. Thiselton, A. Hardcastle, G. Rappold and the Resource Centre of the German Human Genome Project (RZPD) for the provision of clones and mapping data; the HUGO Gene Nomenclature Committee (S. Povey (chair), M. W. Wright, M. J. Lush, R. C. Lovering, V. K. Khodiyar, H. M. Wain and C. C. Talbot Jr) for assigning official gene symbols; C. Rees for assistance with the manuscript; C. Tyler-Smith for critical reading of the manuscript; and the Wellcome Trust, the NHGRI, and the Ministry of Education and Research (Germany) for financial support.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.T.R. (mtr@sanger.ac.uk). All DNA sequences reported in this study have been deposited in the EMBL or GenBank databases, and accession numbers are given in Supplementary Fig. 1.

Mark T. Ross¹, Darren V. Grafham¹, Alison J. Coffey¹, Steven Scherer², Kirsten McLay¹, Donna Muzny², Matthias Platzer³, Gareth R. Howell¹, Christine Burrows¹, Christine P. Bird¹, Adam Frankish¹, Frances L. Lovell¹, Kevin L. Howe¹, Jennifer L. Ashurst¹, Robert S. Fulton⁴, Ralf Sudbrak^{5,6}, Gaiping Wen³, Matthew C. Jones¹, Matthew E. Hurles¹, T. Daniel Andrews¹, Carol E. Scott¹, Stephen Searle¹, Juliane Ramser⁷, Adam Whittaker¹, Rebecca Deadman¹, Nigel P. Carter¹, Sarah E. Hunt¹, Rui Chen², Andrew Cree², Preethi Gunaratne², Paul Havlak², Anne Hodgson², Michael L. Metzker², Stephen Richards², Graham Scott², David Steffen², Erica Sodergren², David A. Wheeler², Kim C. Worley², Rachael Ainscough¹, Kerrie D. Ambrose¹, M. Ali Ansari-Lari², Swaroop Aradhya², Robert I. S. Ashwell¹, Anne K. Babbage¹, Claire L. Bagguley¹, Andrea Ballabio², Ruby Banerjee¹, Gary E. Barker¹, Karen F. Barlow¹, Ian P. Barrett¹, Karen N. Bates¹, David M. Beare¹, Helen Beasley¹, Oliver Beasley¹, Alfred Beck⁵, Graeme Bethel¹, Karin Blechschmidt³, Nicola Brady¹, Sarah Bray-Allen¹, Anne M. Bridgeman¹, Andrew J. Brown¹, Mary J. Brown², David Bonnini², Elspeth A. Bruford⁸, Christian Buhay², Paula Burch², Deborah Burford¹, Joanne Burgess¹, Wayne Burrill¹, John Burton¹, Jackie M. Bye¹, Carol Carder¹, Laura Carrel⁹, Joseph Chako², Joanne C. Chapman¹, Dean Chavez², Ellison Chen¹⁰, Guan Chen², Yuan Chen¹¹, Zhijian Chen², Craig Chinault², Alfredo Ciccodicola¹², Sue Y. Clark¹, Graham Clarke¹, Chris M. Clee¹, Sheila Clegg¹, Kerstin Clerc-Blankenburg², Karen Clifford¹, Vicky Copley¹, Charlotte G. Cole¹, Jen S. Conquer¹, Nicole Corby¹, Richard E. Connor¹, Robert David², Joy Davies¹, Clay Davis², John Davis¹, Oliver Delgado², Denise DeShazo², Pawandeep Dhami¹, Yan Ding², Huyen Dinh², Steve Dodsworth¹, Heather Draper², Shannon Dugan-Rocha², Andrew Dunham¹, Matthew Dunn¹, K. James Durbin², Ireena Dutta¹, Tamsin Eades¹, Matthew Ellwood¹, Alexandra Emery-Cohen², Helen Errington¹, Kathryn L. Evans¹³, Louisa Faulkner¹, Fiona Francis¹⁴, John Frankland¹, Audrey E. Fraser¹, Petra Galgoczy³, James Gilbert¹, Rachel Gill², Gernot Glöckner³, Simon G. Gregory¹, Susan Gribble¹, Coline Griffiths¹, Russell Grocock¹, Yanghong Gu², Rhian Gwilliam¹, Cerissa Hamilton², Elizabeth A. Hart¹, Alicia Hawes², Paul D. Heath¹, Katja Heitmann⁵, Steffen Hennig⁵, Judith Hernandez², Bernd Hinzmann³, Sarah Ho¹, Michael Hoffs¹, Phillip J. Howden¹, Elizabeth J. Huckle¹, Jennifer Hume², Paul J. Hunt¹, Adrienne R. Hunt¹, Judith Isherwood¹, Leni Jacob², David Johnson¹, Sally Jones², Pieter J. de Jong¹⁵, Shirin S. Joseph¹, Stephen Keenan¹, Susan Kelly², Joanne K. Kershaw¹, Ziad Khan², Petra Kioschis¹⁶, Sven Klages⁵, Andrew J. Knights¹, Anna Kosiura⁵, Christie Kovar-Smith², Gavin K. Laird¹, Cordelia Langford¹, Stephanie Lawlor¹, Margaret Leversha¹, Lora Lewis², Wen Liu², Christine Lloyd¹, David M. Lloyd¹, Hermela Louiseged², Jane E. Loveland¹, Jamieson D. Lovell¹, Ryan Lozado², Jing Lu², Rachael Lyne¹, Jie Ma², Manjula Maheshwari², Lucy H. Matthews¹, Jennifer McDowall¹, Stuart McLaren¹, Amanda McMurray¹, Patrick Meidi¹, Thomas Meitinger¹⁷, Sarah Milne¹, George Miner², Shailesh L. Mistry¹, Margaret Morgan², Sidney Morris², Ines Müller^{5,18}, James C. Mullikin¹⁹, Ngoc Nguyen², Gabriele Nordsiek³, Gerald Nyakatura³, Christopher N. O'Dell¹, Geoffery Okwuonu², Sophie Palmer¹, Richard Pandian¹, David Parker², Julia Parrish², Shiran Pasternak², Dina Patel¹, Alex V. Pearce¹, Danita M. Pearson¹, Sarah E. Pelan¹, Lesette Perez², Keith M. Porter¹, Yvonne Ramsey¹, Kathrin Reichwald³, Susan Rhodes¹, Kerry A. Ridler¹, David Schlessinger²⁰, Mary G. Schueler¹⁹, Harminder K. Sehra¹, Charles Shaw-Smith¹, Hua Shen², Elizabeth M. Sheridan¹, Ratna Shownkeen¹, Carl D. Skuce¹, Michelle L. Smith¹,

Elizabeth C. Sotheran¹, Helen E. Steingruber¹, Charles A. Steward¹, Roy Storey¹, R. Mark Swann¹, David Swarbreck¹, Paul E. Tabor², Stefan Taudien³, Tineace Taylor², Brian Teague², Karen Thomas¹, Andrea Thorpe¹, Kirsten Timms², Alan Tracey¹, Steve Trevanion¹, Anthony C. Tromans¹, Michele d'Urso¹², Daniel Verduzco², Donna Villasana², Lenee Waldron², Melanie Wall¹, Qiaoyan Wang², James Warren², Georgina L. Warry¹, Xuehong Wei², Anthony West¹, Siobhan L. Whitehead¹, Mathew N. Whiteley¹, Jane E. Wilkinson¹, David L. Willey¹, Gabrielle Williams², Leanne Williams¹, Angela Williamson², Helen Williamson¹, Laurens Wilming¹, Rebecca L. Woodmansey¹, Paul W. Wray¹, Jennifer Yen², Jingkun Zhang², Jianling Zhou², Huda Zoghbi², Sara Zorilla², David Buck¹, Richard Reinhardt⁵, Annemarie Poustka¹⁶, André Rosenthal³, Hans Lehrach⁵, Alfons Meindl⁷, Patrick J. Minx⁴, LaDeana W. Hillier⁴, Huntington F. Willard²¹, Richard K. Wilson⁴, Robert H. Waterston⁴, Catherine M. Rice¹, Mark Vaudin¹, Alan Coulson¹, David L. Nelson², George Weinstock², John E. Sulston¹, Richard Durbin¹, Tim Hubbard¹, Richard A. Gibbs², Stephan Beck¹, Jane Rogers¹ & David R. Bentley¹

Affiliations for authors: 1, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK; 2, Baylor College of Medicine Human Genome Sequencing Center, Department of Molecular and Human Genetics, One Baylor Plaza, Houston, Texas 77030, USA; 3, Genomanalyse, Institut für Molekulare Biotechnologie, Beutenbergstr. 11, 07745 Jena, Germany; 4, Washington University Genome Sequencing Center, Box 8501, 4444 Forest Park Avenue, St. Louis, Missouri 63108, USA; 5, Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany; 6, Institute for Clinical Molecular Biology, Christian-Albrechts-University, 24105 Kiel, Germany; 7, Medizinische Genetik, Ludwig-Maximilian-Universität, Goethestr. 29, 80336 München, Germany; 8, HUGO Gene Nomenclature Committee, The Galton Laboratory, Department of Biology, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK; 9, Department of Biochemistry and Molecular Biology, Pennsylvania State College of Medicine, Hershey, Pennsylvania 17033, USA; 10, Advanced Center for Genetic Technology, PE-Applied Biosystems, Foster City, California 94404, USA; 11, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK; 12, Institute of Genetics and Biophysics, Adriano Buzzati-Traverso, Via Marconi 12, 80100 Naples, Italy; 13, Medical Genetics Section, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, UK; 14, Laboratoire de Génétique et de Physiopathologie des Retards Mentaux, Institut Cochin. Inserm U567, Université Paris V., 24 rue du Faubourg Saint Jacques, 75014 Paris, France; 15, BACPAC Resources, Children's Hospital Oakland Research Institute, 747 52nd Street, Oakland, California 94609, USA; 16, Molekulare Genomanalyse, Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 580, 69120 Heidelberg, Germany; 17, Institute of Human Genetics, GSF National Research Center for Environment and Health, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany; 18, RZPD Resource Center for Genome Research, 14059 Berlin, Germany; 19, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA; 20, Laboratory of Genetics, National Institute on Aging, 333 Cassell Drive, Baltimore, Maryland 21224, USA; 21, Institute for Genome Sciences & Policy, Duke University, Durham, North Carolina 27708, USA

Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals

Xiaohui Xie¹, Jun Lu¹, E. J. Kulbokas¹, Todd R. Golub¹, Vamsi Mootha¹, Kerstin Lindblad-Toh¹, Eric S. Lander^{1,2*} & Manolis Kellis^{1,3*}

¹Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02141, USA

²Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02139, USA

³Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

* These authors contributed equally to this work

Comprehensive identification of all functional elements encoded in the human genome is a fundamental need in biomedical research. Here, we present a comparative analysis of the human, mouse, rat and dog genomes to create a systematic catalogue of common regulatory motifs in promoters and 3' untranslated regions (3' UTRs). The promoter analysis yields 174 candidate motifs, including most previously known transcription-factor binding sites and 105 new motifs. The 3' UTR analysis yields 106 motifs likely to be involved in post-transcriptional regulation. Nearly one-half are associated with microRNAs (miRNAs), leading to the discovery of many new miRNA genes and their likely target genes. Our results suggest that previous estimates of the number of human miRNA genes were low, and that miRNAs regulate at least 20% of human genes. The overall results provide a systematic view of gene regulation in the human, which will be refined as additional mammalian genomes become available.

Comparative genomics provides a powerful approach for the systematic discovery of functional elements in the human genome^{1–8}, by virtue of their evolutionary conservation across related species. Recent studies have revealed that most functional elements are non-protein-coding; these are likely to include regulatory signals, RNA genes and structural elements. Although the largest and most conserved elements are readily identified⁹, the vast majority of non-coding functional elements remain unknown. It has been particularly difficult to recognize short regulatory sequences, such as the DNA binding sites for transcription factors.

Our goal here is to create a catalogue of 'common regulatory motifs', by which we mean short, functional sequences (typically, 6–10 bases) that are used many times in a genome. Although reliably recognizing each individual occurrence of a motif requires evolutionary comparison with many species¹⁰, it should be possible to define the motifs themselves with many fewer species based on their multiple conserved occurrences in the genome. Recent studies of four related yeast species have shown the potential power of this approach in a relatively small genome^{11,12}.

To apply this approach to the vastly larger human genome, we focused on two limited subsets that are likely to be highly enriched for regulatory elements: promoter regions and 3' UTRs of protein-coding genes. We aligned these regions across the human, mouse, rat and dog genomes^{13–15}, and searched for highly conserved, frequently occurring sequence patterns. In promoter regions, we discovered 174 candidate motifs, including most previously known transcription-factor binding sites and 105 new motifs that seem to be functional based on multiple criteria. In 3' UTRs, we discovered 106 motifs likely to be involved in post-transcriptional regulation. Nearly one-half of these are associated with miRNAs, demonstrating the extraordinary importance of this recently discovered regulatory mechanism¹⁶. Our results suggest that previous estimates of human miRNA number were low¹⁷, and that at least 20% of human genes are probably regulated by miRNAs. More broadly, our results suggest that it will be possible to construct a comprehensive catalogue of common regulatory motifs in the human genome as more species become available.

Alignment of promoters and 3' UTRs

We first constructed genome-wide alignments for the four mammalian species in promoter regions and 3' UTRs. We studied a total of ~17,700 well-annotated genes from the RefSeq database¹⁸, with a single reference messenger RNA selected for each gene. Each promoter region was defined as the non-coding sequence contained within a 4-kilobase (kb) window centred at the annotated transcriptional start site (TSS), and each 3' UTR was defined based on the annotation of the reference mRNA (see Methods). The promoter set contains ~68 megabases (Mb) of sequence, whereas the 3' UTR set comprises ~15 Mb. In addition, we defined a control set comprising ~123 Mb of intronic sequence consisting of the last two introns from the genes; these sequences were used as a control because terminal introns are thought to contain relatively fewer regulatory elements.

The human sequence in each set was aligned to the other three mammalian genomes^{19,20}. Some segments could not be aligned in all four species, primarily because they represent new sequence recently inserted by transposons, ancestral sequence deleted in one of the other species, or regions still missing from the draft genome sequences (see Supplementary Information). The proportion of bases that could be aligned across all four species was ~51% for promoters (44% upstream and 58% downstream of the TSS) and ~73% for the 3' UTRs. These proportions are much higher than for the control intronic regions (34%) or for the whole genome (28%), presumably reflecting the presence of important conserved elements (see Supplementary Information).

The total number of aligned bases is ~35 Mb for the promoters and ~11 Mb for the 3' UTRs, which is comparable to the size of yeast genomes. Additionally, the evolutionary tree (Fig. 1a) of the four mammals has a total branch length similar to that used for comparative analysis of *Saccharomyces* species¹¹ (0.68 substitutions per base in human promoters versus 0.83 in yeast intergenic regions; see Supplementary Information).

Conservation properties of regulatory motifs

Our comparative analysis can be illustrated by considering a known

regulatory motif. The 8-mer TGACCTTG is known to be a binding site of the Err- α protein and to occur in the promoters of many genes induced during mitochondrial biogenesis^{21,22}. The promoter of the *GABPA* gene contains a well-studied Err- α -binding site, which is conserved across all four species, and stands out from the non-conserved flanking sequence (Fig. 1b). More generally, the Err- α motif occurs 434 times in human promoter regions and 162 of these occurrences are conserved across all four species; the conservation rate is thus 37%. By contrast, a random 8-mer motif shows a markedly lower conservation rate in promoters (6.8%). Moreover, the high conservation of the Err- α motif is specific to promoter regions; it shows a much lower conservation rate in introns (6.2%).

We quantified the extent of excess conservation by using a motif conservation score (MCS), which essentially represents the number of standard deviations (s.d.) by which the observed conservation rate of a motif exceeds the expected conservation rate for comparable random motifs (see Methods). A 'conserved occurrence' of a (possibly degenerate) motif is an instance in which an exact match to the motif is found in all four species; the 'conservation rate' of a motif is defined to be the ratio of conserved occurrences to total occurrences in the human genome. Comparable random motifs are generated by sampling human genomic sequence in promoters (divided according to high or low CpG content) or 3' UTRs in order to ensure equivalent sequence composition (see Methods).

For the Err- α -binding motif, the MCS is 25.2 s.d. (the binomial probability of observing 162 conserved instances out of 434, given an expected rate of conservation of 6.8%).

We examined the conservation properties of known motifs in mammalian promoters using the TRANSFAC database of reported transcription-factor binding sites²³. The 446 motifs in TRANSFAC were clustered on the basis of sequence similarity, resulting in 123 motif clusters (see Methods). Most of these 123 motifs had high MCS, with 63% having MCS > 3 and nearly one-half having MCS > 5. By contrast, a control set of random motifs has only 1.6% with MCS > 3 and none with MCS above 5. The TRANSFAC motifs with low conservation scores have three likely explanations: they may be erroneous or over-specified (there is no uniform standard of evidence for inclusion in the database), they may have diverged between the species compared, or they may have too few biologically functional occurrences for the conserved instances to stand out with sufficient statistical significance.

For 3' UTRs, few common motifs have been defined and there is no database analogous to TRANSFAC. One well-known example is the polyadenylation signal (AATAAA). This motif indeed shows a high conservation rate, with 6,617 out of 14,266 (46%) occurrences in 3' UTRs being conserved (conservation rate = 46%, MCS = 135). Control random motifs show a conservation rate of only 10% in 3' UTRs, and the polyadenylation signal shows a conservation rate of only 6% in the intronic controls.

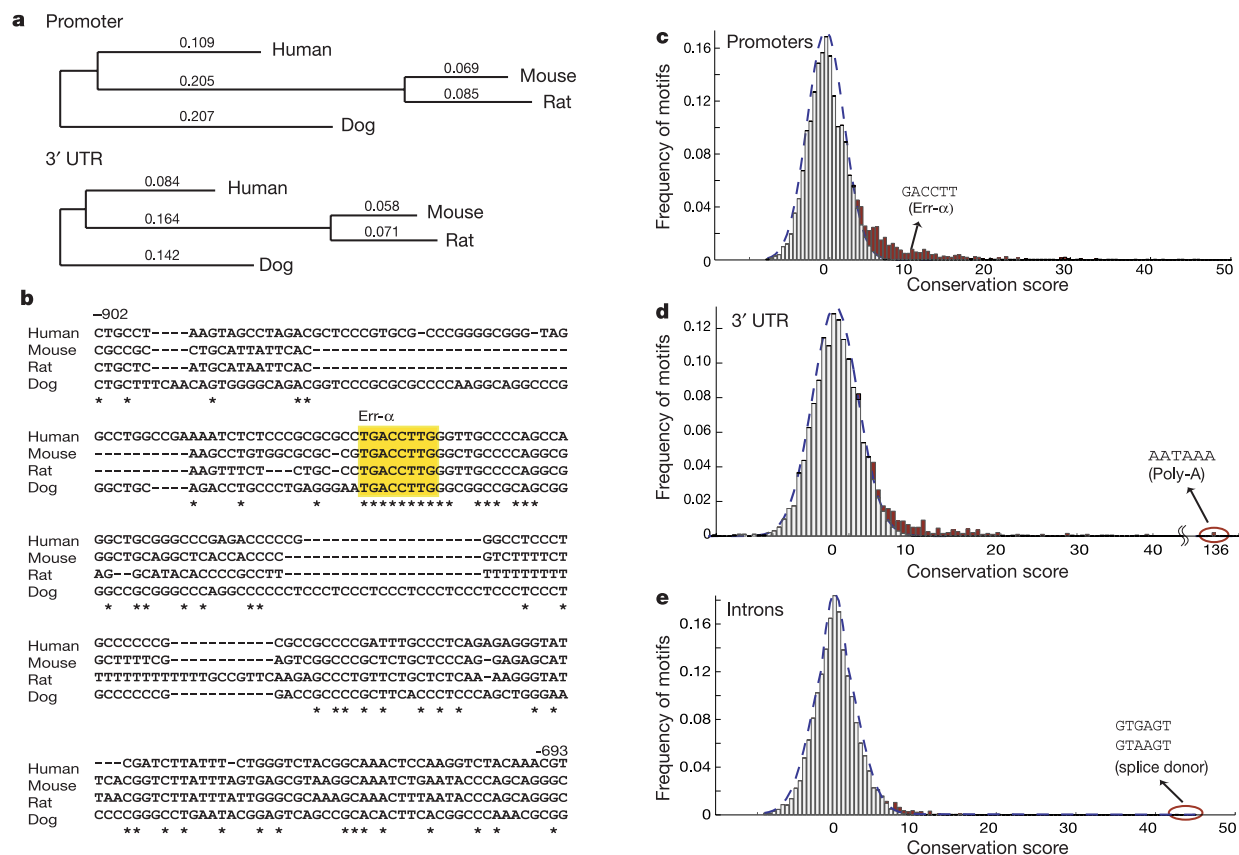


Figure 1 Conservation properties in human promoter regions and 3' UTRs.

a, Evolutionary tree relating the four mammalian species. Branch lengths denote number of substitutions per site. Average nucleotide per cent identity to human is 62% for mouse, 60% for rat and 69% for dog in promoter regions, and respectively 68%, 67% and 76% in 3' UTRs. **b**, Conservation in *GABPA* promoter region reveals functional Err- α motif. Asterisks denote conserved bases. The yellow box marks the experimentally validated

Err- α -binding site. **c-e**, Excess conservation in promoter and 3'-UTR regions reveals short sequences under evolutionary selection. Motif conservation score (MCS) distribution is shown for all 6-mer motifs in aligned promoters (**c**), 3'-UTR regions (**d**) and introns (**e**). The dashed curve shows fit to gaussian distribution. Excess conservation relative to this distribution is shown in red.

Discovery of new motifs

We next explored the general conservation properties of short sequences within promoters and 3' UTRs. Specifically, we measured the conservation rate of all 6-mer sequences. For both regions, there is a large excess of 6-mers with high MCS scores (Fig. 1c, d). The distribution of MCS scores suggests that a significant proportion of 6-mers are under purifying selection (13% in promoters and 8.6% in 3' UTRs, indicated by the shaded regions in Fig. 1c, d). By contrast, intronic regions have a largely symmetric distribution (Fig. 1e) with only a modest number of outliers. (The two most extreme cases correspond to known mRNA splice donor signals, GTGAGT and GTAAGT.)

On the basis of these properties, we set out to systematically discover common regulatory motifs in promoters and 3' UTRs. We evaluated motifs containing 6–18 bases, with degeneracy allowed at each position; a computational algorithm was developed to make it feasible to screen $\sim 10^{12}$ possible motifs across the genomic regions (see Methods). Motifs with high conservation scores (MCS > 6) were identified, similar motifs were clustered together and the motif with the highest MCS in each cluster was selected (see Methods). This resulted in a small number of motifs with MCS > 6, which we refer to as 'highly conserved motifs'

Motifs in promoters

In promoters, we found 174 highly conserved motifs. Among these, some could be immediately recognized as known regulatory elements. For example, the three strongest motifs correspond to binding sites for the oncogene regulator ELK-1, the cell-cycle regulator Myc and the mitochondrial respiratory chain regulator NRF-1. Overall, 59 discovered motifs showed strong matches to known motifs and 10 showed weaker matches (see Supplementary Information), together accounting for 72% of the 123 previously known motifs that we assembled from the TRANSFAC database (Supplementary Information). (By contrast, a comparable collection of random motifs would match only $\sim 5\%$ of the TRANSFAC motifs; see Supplementary Information.)

The remaining 105 discovered motifs represent potentially new regulatory elements. The list includes numerous notable examples, some with conservation scores much higher than for most known motifs. For example, the newly discovered motif M₄ (ACTAYRNNCCCR) occurs 520 times in human promoter regions, of which 317 (61%) are conserved, and the new near-palindromic motif M₈ (TMTGCGGANR) occurs 368 times, of which 236 (64%) are conserved.

In the absence of specific information about the role of these

Table 1 Top 50 of 174 discovered motifs in human promoters

No.	Discovered motif	MCS	Known factor*	Conservation rate†	Tissue enrichment‡	Position bias§
1	RCGCAnGCGY	107.8	NRF-1	0.49	15.0	−62
2	CACGTG	85.3	MYC	0.47	8.8	−62
3	SCGGAAGY	80.4	ELK-1	0.44	22.4	−24
4	ACTAYRnnnCCCR	69.5	—	0.61	8.1	−89
5	GATTGGY	64.6	NF-Y	0.51	9.8	−63
6	GGGCGGR	63.9	SP1	0.21	11.4	−63
7	TGAnTCA	62.8	AP-1	0.38	6.5	—
8	TMTGCGGAnR	55.7	—	0.64	9.4	−62
9	TGAYRTCA	55.7	ATF3	0.50	6.1	−66
10	GCCATnTTG	54.7	YY1	0.72	12.2	—
11	MGGAAAGTG	51.6	GABP	0.43	13.9	−23
12	CAGGTG	47.6	E12	0.26	9.9	—
13	CTTTGT	46.0	LEF1	0.42	13.6	—
14	TGACGTCA	44.8	ATF3	0.44	4.2	−22
15	CAGCTG	43.9	AP-4	0.27	8.9	—
16	RYTTCCTG	43.0	C-ETS-2	0.32	7.4	−24
17	AACTTT	42.1	IRF1	0.43	11.1	—
18	TCAAnTGAY	40.4	SREBP-1	0.47	4.9	−64
19	GKCGCn(7)TGAYG	40.1	—	0.35	5.6	−62
20	GTGACGY	38.4	E4F1	0.34	6.6	−56
21	GGAAnCGGAAnY	37.7	—	0.68	7.0	−33
22	TGCGCAnK	37.4	—	0.24	8.2	−17
23	TAATTA	37.3	CHX10	0.29	7.1	—
24	GGGAGGRR	33.5	MAZ	0.16	9.4	—
25	TGACCTY	33.4	ESRRA	0.30	7.7	—
26	TTAYRTAA	32.6	E4BP4	0.34	6.1	—
27	TGn(6)KCCAR	32.3	—	0.27	4.5	—
28	CTAWWWATA	32.3	RSRFC4	0.36	7.6	—
29	CTTTAAR	30.8	—	0.43	5.4	—
30	YCGGYRCGC	30.5	—	0.19	5.2	−31
31	GGGYGTGnY	30.0	—	0.24	5.4	−63
32	TGASTMAGC	27.2	NF-E2	0.39	5.4	−66
33	YTATTTnR	26.4	MEF-2	0.21	7.1	—
34	CYTAGCAAY	26.1	—	0.50	5.2	−142
35	GCAAnCTGnY	25.7	MYOD	0.25	8.2	—
36	RTAAACA	25.6	FREAC-2	0.46	7.0	—
37	GTTRYCATRR	25.3	—	0.54	7.6	−56
38	TGACCTTG	25.2	ERR-α	0.37	8.1	—
39	TCCCRnnRTGC	24.3	—	0.30	6.8	−60
40	TTCYnRGAA	24.3	STAT5A	0.19	—	—
41	TGACAGnY	24.1	MEIS1	0.27	6.9	—
42	TGACATY	23.8	—	0.23	5.8	—
43	GTTGnYnnRGnAAC	23.7	—	0.47	4.7	−57
44	YATGnWAAT	23.5	OCT-X	0.53	6.9	—
45	CCAnnAGRKGCC	23.4	—	0.47	—	−101
46	WTTGKCTG	23.0	—	0.25	5.0	−63
47	TGCCAAR	22.9	NF-1	0.25	7.0	—
48	GCGnAnTTCC	22.8	C-REL	0.30	6.0	−12
49	CATTGTYY	22.5	SOX-9	0.43	5.8	—
50	RGAGGAARY	22.4	PU.1	0.22	4.0	—

*Name of the best-matching motif in TRANSFAC database, if any. Weak matches are indicated in *italics*.

†The percentage of human motif occurrences that match the motif consensus across all four species.

‡A measure of the maximum enrichment of conserved motif occurrences upstream of genes expressed in a compendium of 75 human tissues.

§For motifs with strong positional bias (score > 5 s.d.), the mode of the distance distribution upstream of the TSS.

putative motifs, we used two approaches to demonstrate that most of the new motifs are likely to be biologically meaningful. First, we correlated the presence of motifs with the tissue specificity of gene expression. We reasoned that the genes controlled by a common regulator would often (although not always) show enriched expression in specific sets of tissues. For each motif, we defined the set S_1 of genes with conserved occurrences of the motif and an equal-sized control set S_2 of genes in which the motif occurs in the human genome but is not conserved (see Methods). Using gene expression data from 75 human tissues²⁴, significant enrichment in one or more tissues (z -score > 4.0) was seen for 59 of the 69 (86%) known motifs and 53 of the 105 (50%) new motifs (Table 1 and Fig. 2). In contrast, the control sets show little or no enrichment across the same tissues (Supplementary Fig. S2). For example, the best-conserved new motifs M_4 and M_8 show enrichment in haematopoietic cells, motif M_{27} (TGGN₆KCCAR) is enriched in trachea and lung, and motif M_{29} (CTTTAAR) is enriched in brain-related tissues such as pons, parietal lobe and cingulate cortex.

Second, we examined positional bias of the motifs relative to the TSS. Although the analysis covered a 4-kb region surrounding the TSS, the discovered motifs preferentially occur in the human genome within ~ 100 bases of the TSS, and their conserved occurrences across all four species show an even more marked

enrichment in this region (Fig. 3a, b), consistent with the motifs being involved in transcription initiation. Within this overall trend, certain motifs show distinctive positional preferences. Approximately 28% of the known motifs and 35% of the new motifs show significant positional preference (see Methods). For example, the two strongest new motifs (M_4 and M_8) tend to occur at distances centred around -89 and -62 bp upstream of the TSS, respectively (Table 1 and Fig. 3c, d). Overall, 89% of the known motifs and 69% of the new motifs show tissue specificity, positional bias or both. Taken together, these results strongly suggest that the new promoter motifs are likely to be biologically meaningful.

In addition, several of the motifs tend to appear in multiple copies in the promoter. For example, 17.5% of genes containing motif M_4 have more than one copy of M_4 within 200 bp of each other (27-fold increase compared with 0.66% expected at random), and 10% of genes containing motif M_8 have multiple copies of this motif within 200 bp (compared with 0.26% expected by chance). Such clustering is a common feature of several known regulatory motifs.

Motifs in 3' UTRs

We next turned to motif discovery in 3'-UTR regions²⁵. Using the same approach as for promoter regions, we discovered 106 highly

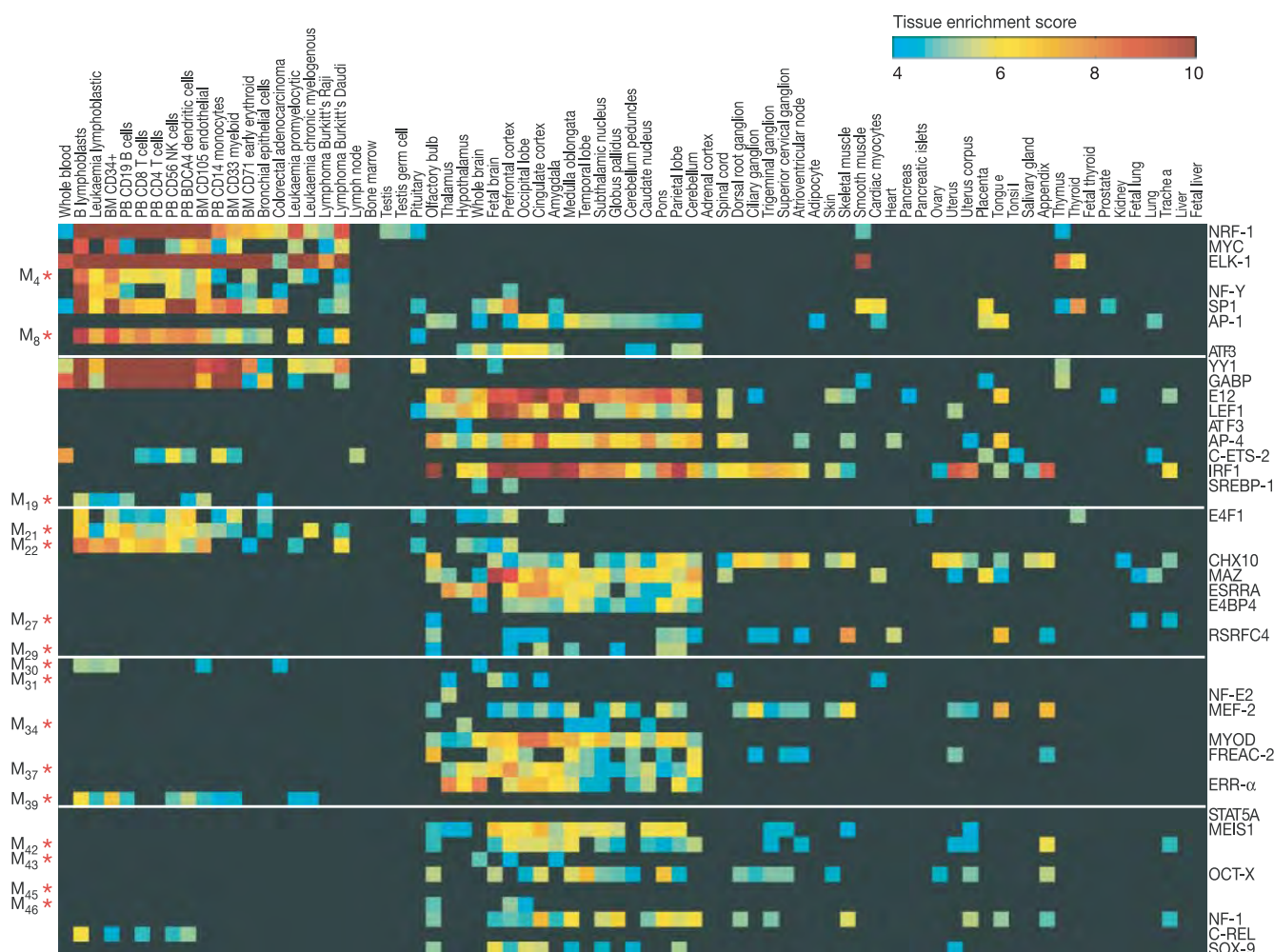


Figure 2 Tissue specificity of expression for genes containing discovered motifs. For each of the 174 motifs, we defined the set of genes whose promoters contain conserved occurrences of the motif, and tested for enriched expression in 75 human tissues. The enrichment score (see Methods) is represented in pseudo-colour, with only scores greater than 4 shown. Motifs are ordered by MCS; asterisks denote new motifs (left); factor names

are shown for known motifs (right). Only the top 50 motifs are shown. The maximum enrichment score across all tissues is reported in Table 1. Control gene sets were also constructed for each motif, consisting of an equal-sized set of genes in which the motif occurs but is not conserved; these control sets show little or no enrichment across the same tissues (see Supplementary Fig. S2).

conserved motifs in 3' UTRs ($MCS > 6$). Because 3'-UTR motifs have not been extensively studied, we could not compare the discovered motifs to a large collection of previously known motifs; however, the motifs show two unusual properties that give insight into their nature.

First, the 3'-UTR motifs show a strong directional bias with respect to DNA strand. Whereas promoter motifs tend to have similar conservation rates on the coding and non-coding strands, the 3'-UTR motifs are preferentially conserved in only one strand (Fig. 4a). The strand specificity is consistent with the 3'-UTR motifs acting at the level of RNA rather than DNA and thus having a role in post-transcriptional regulation.

Second, the 3'-UTR motifs show an unusual length distribution. They have a strong peak at an 8-base length, whereas no such bias was found for promoter motifs (Fig. 4b). Moreover, the motifs of length 8 have a strong tendency to end with the nucleotide 'A' (Fig. 4c). These properties are reminiscent of a feature of miRNA¹⁶: many mature miRNAs start with a 'U' base followed by a 7-base 'seed' complementary to a site in the 3' UTR of target mRNAs^{17,26,27}; binding at such sites can guide degradation or repress translation of the mRNA. We reasoned that many of the highly conserved 8-mer motifs discovered by our unbiased procedure might be binding sites for conserved miRNAs.

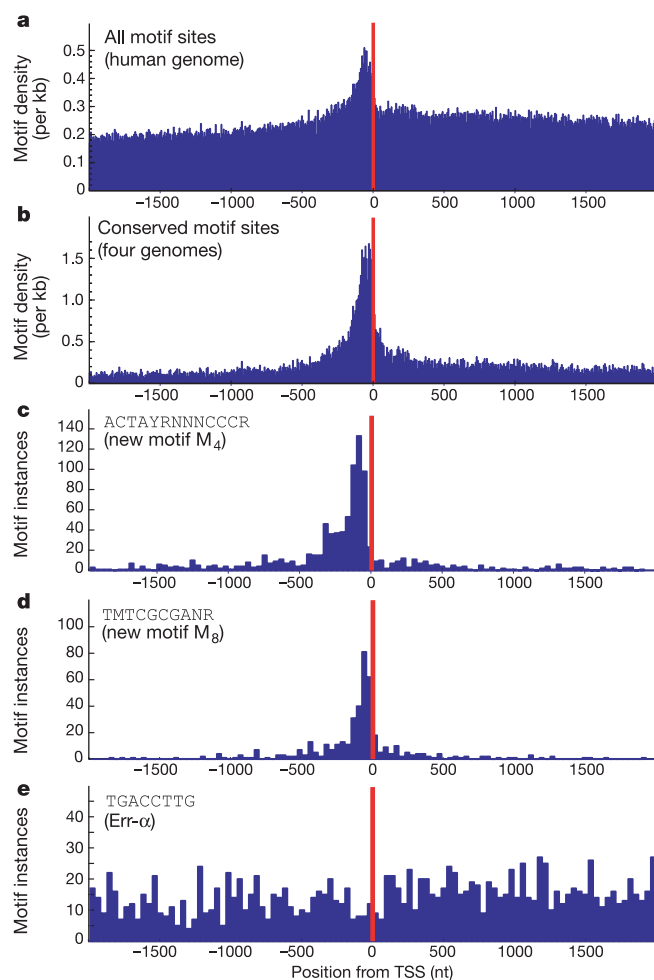


Figure 3 Discovered promoter motifs show positional bias with respect to transcriptional start site (TSS). **a**, Distribution of distance from TSS for all occurrences in human genome peaks within 100 bp before TSS. **b**, Distribution for conserved occurrences shows an even stronger peak. **c, d**, New motifs M_4 and M_8 peak at -81 and -69 respectively. **e**, Some motifs do not show specific peaks, including the known $Err-\alpha$ motif. nt, nucleotides.

Relationship with miRNAs

To investigate the relationship with miRNAs, the motif discovery procedure was repeated using only contiguous, non-degenerate 8-mers. We identified the subset of 8-mers with conservation rate $> 18\%$ (compared with the rate for a random 8-mer of 7.6%) and clustered these 8-mers into motifs defined as sets of similar 8-mers (see Supplementary Information). We refer to these as 'highly conserved 8-mer motifs'. We obtained 72 highly conserved 8-mer motifs, which match 46% of the full set of 3'-UTR motifs (Fig. 4b).

We then searched for complementary matches of the 8-mer motifs to the 207 distinct human miRNAs (encoded by 222 miRNA genes) listed in the current registry²⁸. Roughly 43.5% of the known miRNAs can match through Watson–Crick pairing to the highly conserved 8-mer motifs (versus 2% for an equal number of random 8-mers). When we relax the Watson–Crick base pairing to allow one mismatch, an additional 27 miRNAs can be identified (including four miRNAs containing T–G pairing and ten miRNAs whose first 5' base was paired unselectively to 'A' in the conserved 8-mers). Moreover, the matches begin at nucleotide 1 or 2 of the miRNA gene in more than 95% of cases (Fig. 4d). These results strongly suggest that these 8-mer motifs represent target sites for miRNAs.

The remainder of the known miRNA genes do not contain complementary matches to any of the highly conserved 8-mer motifs. These miRNAs may bind relatively few targets, form only partial complements with their targets, tend to bind regions outside the 3' UTR, or have been diverged between the species we are comparing. It is worth noting that, among the known miRNAs, those that do not match highly conserved 8-mers are evolving much more rapidly; they show a rate of disruptive mutation that is fivefold higher than for known miRNAs that match the highly conserved 8-mers. This suggests that the miRNA genes that match the highly conserved 8-mers have many more targets and, as a result, are much more constrained in their evolution.

New miRNA genes

We then sought to use the 8-mer motifs to discover new miRNA genes. Towards this end, we searched the four aligned genomes for conserved sequences complementary to any of the 72 discovered 8-mer motifs, extracted the sequence flanking each conserved site, and used the published RNAfold^{29,30} program to test for a conserved RNA-folding pattern characteristic of miRNA genes (stem-loop structures with calculated folding free energy of at least 25 kcal mol^{-1} in all four species).

We identified 242 conserved and stable stem-loop sequences, containing conserved instances of most (52 of 72) of the highly conserved 8-mer motifs (see Supplementary Information). These include 113 sequences encoding known miRNA genes (52% of 222 known miRNA genes) and 129 sequences encoding predicted new miRNAs (Table 2 and Fig. 4e; see also Supplementary Table S8). The predicted miRNA genes show relatively few mutations that would disrupt an inferred base pair in the stem structure; the frequency of such mutations is similar to that seen for the 222 known miRNA genes.

A representative set of 12 predicted new miRNA genes was selected for validation according to accepted criteria^{16,31} (involving purification of small RNAs, ligation of adaptors, polymerase chain reaction (PCR) amplification, cloning and DNA sequencing to verify the precise sequence and junction; see Methods). Lacking prior information about tissue or temporal specificity of expression, we used pooled small RNAs prepared from ten human tissues (breast, pancreas, prostate, colon, stomach, uterus, lung, brain, liver and kidney) (see Methods). This may miss many miRNAs expressed primarily during development or at low levels in the adult, or not expressed in the tissues we tested. Nonetheless, 6 of the

12 (50%) predicted miRNA genes were found to be clearly expressed in the pooled adult tissues.

We conclude that many of the 129 candidates are likely to be bona fide miRNA genes. Moreover, there are likely to be many additional miRNAs inasmuch as our approach was only designed to find those containing a conserved 8-mer and thus only detects one-half of known miRNAs. This challenges a recent conclusion, based on

computational analysis, that nearly all human miRNA genes have already been discovered, and that fewer than 40 remain to be found²⁷.

Targets of miRNAs

The properties of the conserved motifs also allow inferences about the prevalence of regulation by miRNAs. Roughly 40% of human 3'

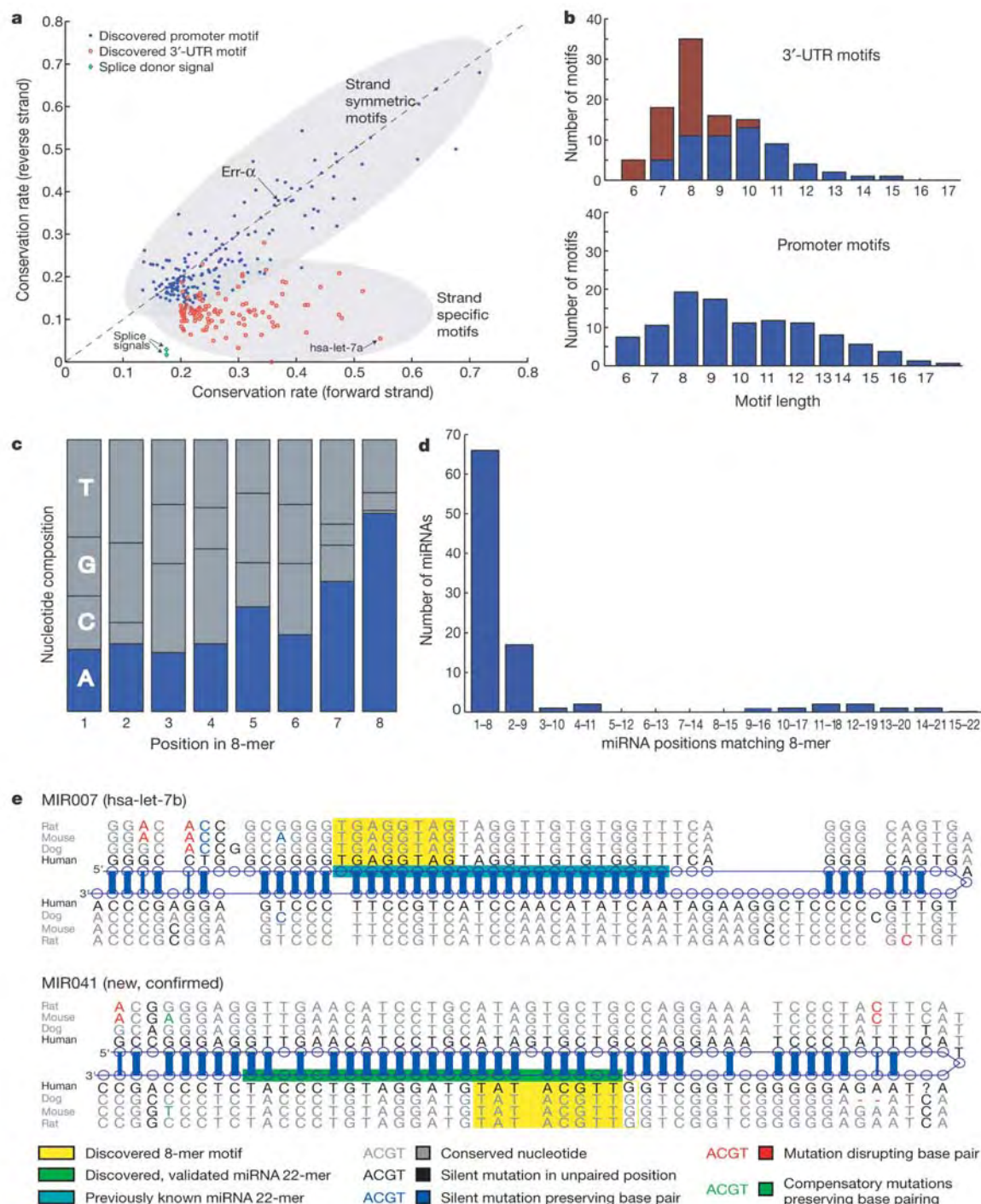


Figure 4 Properties of discovered 3'-UTR motifs and corresponding miRNA genes. **a**, Directionality of 3'-UTR motifs revealed by comparing conservation on forward and reverse strands. Strand preference is also seen for splice signals, but conservation of promoter motifs is largely symmetric. *hsa-let-7a*, *Homo sapiens* let-7a. **b**, Length distribution of 3'-UTR motifs shows abundance of motifs of length 8, but no such preference is seen for promoter motifs. Motifs overlapping conserved 8-mers (red)

account for the length bias. **c**, A total of 72% of 8-mer motifs end with nucleotide A, suggesting complementarity with mature miRNA genes, frequently starting with U. **d**, Ninety-five per cent of discovered 8-mers match known miRNA genes at position 1–8 or 2–9. **e**, Alignments of known and new miRNA stem-loop structures, identified by their complementarity to a discovered 8-mer. New ~22-bp mature miRNA products predicted based on the observation shown in **d** and validated experimentally.

UTRs contain a conserved occurrence of one of the miRNA-associated 8-mer motifs, whereas only ~25% contain conserved occurrences of a comparable control set. This suggests that at least 20% of 3' UTRs may be targets for conserved miRNA-based regulation at the 8-mer motifs (see Methods). This greatly expands previous estimates of the number of targets of miRNA genes¹⁷. With sequence from more mammalian genomes, it should be possible to distinguish the conserved target sites with high specificity and sensitivity¹⁰. There are likely to be additional miRNA target sites in the human genome that are not conserved across all mammals; it

should be possible to find most of these by genomic comparison with closer relatives such as primates.

The results thus provide an unbiased assessment of the relative importance of miRNA-based regulation in the human genome, consistent with recent independent studies^{32–34}. Overall, ~45% of the 108 highly conserved motifs in 3' UTRs appear to be related to miRNA regulation and ~5,000 human genes (~20% of the genome) are likely to be regulated by miRNAs through these conserved motifs.

Remaining motifs in 3' UTRs

When the miRNA-related motifs are removed, there remains a list of 60 3'-UTR motifs with a length distribution similar to that seen for promoters (Fig. 4b). Although there is no systematic procedure for evaluating these motifs, the list contains several interesting features. It includes the known polyadenylation signal and various (A + T)-rich elements³⁵, which are believed to be involved in controlling mRNA stability and degradation. It also contains 11 motifs with a TGTA core sequence, flanked by different variants of an (A + T)-rich element (the most conserved being TGTANATA). Recent work in yeast has identified such motifs as binding sites for the Puf family of RNA-binding proteins; they may represent binding sites for the homologous mammalian proteins (such as PUM1 and PUM2 in human³⁶).

Conclusion

Our comparative genomic analysis of four mammalian species has provided an initial systematic catalogue of human regulatory motifs in promoters and 3' UTRs. In promoters, the approach automatically rediscovered many known motifs and discovered many new ones that appear to be functional based on multiple criteria; many of these show tissue-specific expression and distance constraints with respect to the TSS. In 3' UTRs, the approach provided a first view of common regulatory motifs. Notably, roughly one-half of the discovered motifs in 3' UTRs appear to be related to miRNAs, showing the extraordinary importance of this recently discovered regulatory mechanism. Moreover, these motifs also led to the identification of numerous new miRNA genes. The remaining motifs may be targets of RNA-binding proteins, with a few matching known motifs. The next challenge will be to develop systematic methods to discern the specific functions of these motifs in a genome-wide fashion.

The results here are, of course, only a step towards a comprehensive inventory of human regulatory motifs to serve as a foundation for understanding cellular circuitry and its role in health and disease. Our analysis used stringent thresholds to focus on the most abundant and most conserved motifs. With sequence from a few dozen additional mammals³⁷, it should be possible to create a complete dictionary of such common functional elements. □

Methods

Motif conservation score

Regulatory motifs are represented as consensus sequences (profiles), over an alphabet of 11 characters, consisting of A, C, G, T and N, and the six twofold degenerate characters. A conserved occurrence of a motif *m* matches the motif consensus in each of the four species. The observed conservation rate *p* of a motif is the proportion of human occurrences that are conserved across all four species. The expected conservation rate *p*₀ for a motif of given length and redundancy is computed as the average observed conservation rate of 1,000 random motifs of the same length and redundancy, obtained by sampling the human genome in 1,000 loci. The MCS of a motif *m* is evaluated by comparing its observed conservation rate *p* to the expected conservation rate *p*₀; it corresponds to the binomial probability of observing *k* conserved instances out of total *n* instances given probability *p*₀ of conservation for any one instance.

Motif discovery overview

The MCS is evaluated separately for each type of region (promoters, introns, 3' UTR), thus matching the specific nucleotide composition of each type of region and eliminating biases from the scoring scheme. Additionally, for promoter motifs, the MCS is evaluated separately for sequences inside CpG islands and those outside CpG islands, thus accounting for their radically different nucleotide compositions. In each type of region

Table 2 Top 50 conserved 8-mers in 3' UTRs and corresponding miRNAs

No.	Motif	Conservation rate	miRNA*
1	GTGCAATA	0.55	miR-92, miR-32, miR-137, miR-367, miR-25, miR-217(t), new(12)
2	GTGCCTTA	0.54	miR-124a, miR-224(t), miR-208(t), miR-34b(t), miR-9*(t), miR-34c(t), miR-330(t), new(6)
3	CTACCTCA	0.53	miR-98, let-7i, let-7g, let-7f, let-7e, let-7c, let-7b, let-7a, let-7d, miR-196b, miR-196a, new(4)
4	ACCAAAAGA	0.49	miR-9, new(11)
5	TGTTTACA	0.48	miR-30e-5p, miR-30d, miR-30c, miR-30b, miR-30a-5p, new(4)
6	GCACTTTA	0.48	miR-20, miR-106b, miR-18(t), miR-93, miR-372, miR-17-5p, miR-106a, miR-302d, miR-302c, miR-302b, miR-302a, miR-373, new(4)
7	TGGTGCTA	0.43	miR-29c, miR-29b, miR-29a, miR-107(t), miR-103(t), new(6)
8	CTATGCAA	0.42	miR-153, new(9)
9	TACTTGAA	0.42	miR-26b, miR-26a, new(4)
10	CGCAAAAA	0.42	New(2)
11	GTGCCAAA	0.41	miR-96, miR-182, miR-183, new(16)
12	GTACTGTA	0.40	miR-101, miR-199a±, miR-144, new(2)
13	ATACGGGT	0.40	miR-99a, miR-100, miR-99b(t)
14	AAGCACAA	0.40	miR-218, new(8)
15	TTTGCACT	0.37	miR-19b, miR-19a, miR-301, miR-130b, miR-130a, miR-152, miR-148b, miR-148a, miR-139, new(10)
16	TGTACATA	0.36	–
17	AAGCCATA	0.35	miR-135b, miR-135a
18	ACTGTGAA	0.35	miR-27b, miR-27a, miR-128b, miR-128a, miR-23b(t), miR-23a(t), new(5)
19	AGACAATC	0.33	miR-219, new(2)
20	TGCTGCTA	0.33	miR-195, miR-16, miR-15b, miR-15a, miR-338(t), miR-424, new(5)
21	TTTTGTAC	0.32	New(1)
22	ACATTCCA	0.32	miR-206, miR-1, miR-122a(t)
23	TGAATGTA	0.31	miR-181b(t), miR-181c, miR-181a
24	ACGGTACA	0.30	–
25	CAGTATTA	0.30	miR-200c, miR-200b, new(1)
26	TTGCATGT	0.29	New(7)
27	TCGCATGA	0.29	New(1)
28	CTCAGGGA	0.29	miR-125b, miR-125a, new(6)
29	CAAGTGCC	0.28	New(2)
30	ACTACTGA	0.28	–
31	TGGACCAA	0.28	miR-133b, miR-133a, new(3)
32	GTAATAGT	0.28	New(1)
33	TGTAGATA	0.28	–
34	AACTACTA	0.27	miR-142-3p, new(3)
35	GTACAGTT	0.26	New(1)
36	CACCAGCA	0.26	miR-138(t), new(4)
37	GGTACGAA	0.25	miR-126(t)
38	TGTATAGT	0.24	miR-381
39	AAGGGCTA	0.24	New(1)
40	AGCTTTAA	0.24	New(1)
41	ATTATTCG	0.23	–
42	GGCAGCTA	0.23	miR-22(t), new(1)
43	GCTGTAAA	0.23	New(4)
44	GCACTAAT	0.22	–
45	AAAGGTGC	0.22	–
46	ATGTAGCA	0.22	miR-221(t), miR-222(t)
47	ACACTGGA	0.21	miR-199b(t), miR-199a(t), miR-145(t), new(2)
48	GTATATAG	0.21	–
49	TTTGATAA	0.21	miR-361(t)
50	AAGCATGC	0.21	New(1)

Top 50 of 72 highly conserved 8-mer motifs in 3' UTRs and their corresponding miRNAs.
*Known human miRNAs matching complements of each 8-mer and its variants grouped in the same cluster. The dagger symbol (in parentheses) indicates miRNAs with one mismatch. When multiple new miRNAs are discovered using the 8-mer motifs as seeds, their number is indicated in parentheses. (See Supplementary Information for full alignments of known and discovered miRNAs.)
‡The miRNA mature product comes from the 3' arm of the stem loop.

motifs are enumerated by hashing and refinement of 6-mer seeds possibly with a central gap. All motifs above the cutoff of MCS > 6 are selected, corresponding to specificity values higher than 98%. The selected motifs are grouped into clusters using genome-wide co-occurrence and sequence similarity: we first merge motif pairs which overlap by more than 80% of their sites, keeping only the motif with the highest MCS score; we then cluster the remaining motifs based on their pairwise sequence similarity, defined as the Pearson correlation of their equivalent position weight matrices.

Motif gene-set enrichment score (MGES)

The enrichment of a motif m in a given tissue is evaluated as the enrichment of its gene set S in the ranked list L for that tissue. The non-randomness of the ranks of S in the list L is evaluated using the Mann–Whitney rank sum statistic; the MGES is reported as the standard deviations corresponding to this statistic. For each motif m , three gene sets are generated: a target gene set S_1 , and two control gene sets, S_2 and S_3 , with the same number of genes. The target gene set S_1 of ‘conserved instances’, consists of all genes whose promoters contain at least one conserved instance of the motif m . The control gene set S_2 consists of genes with ‘non-conserved instances’ randomly sampled from genes containing instances of the motif in the human genome, regardless of conservation. The control gene set S_3 consists of genes with ‘shuffled conserved instances’, randomly sampled from the union of all conserved gene sets S_1 , across all motifs.

Identification of new miRNAs

Conserved occurrences of the 3′ UTR 8-mer motifs are identified in the entire human genome by searching both strands for motifs reverse complementary to each 8-mer, and excluding previously annotated functional elements. The neighbourhoods of these alignments are then searched for stable stem-loops over a sliding window of 110 bp every 3 bp, and those with a folding free energy of at least 25 kcal mol^{−1} in each aligned species are selected using the program RNAfold. These were further evaluated based on several observed features of known miRNAs: higher conservation in the ~22 bp stem, lower conservation in the loop and surrounding regions, and appropriate base-pairing and bulges in the stem region.

Experimental validation of new miRNA genes

A set of 12 new miRNA genes was selected for experimental validation, such that their conservation properties and folding free energy values is representative of the set of all 258 predicted miRNAs. The experimental procedure was carefully designed to ensure that the exact ~22 bp predicted product is specifically expressed. First, three steps of rigorous PAGE purification of small RNA fractions exclude large RNA and genomic DNA. Second, PCR is done with one primer specific to the gene, and one complementary to the artificial adaptors used in miRNA ligation. Finally, the resulting clones are sequenced and the precise expected sequence and junction are verified.

Additional methods

Supplementary Information includes a complete description of the methods for whole-genome alignment, definition of promoter and 3′ UTR databases, rapid motif enumeration and scoring, motif clustering, comparison with TRANSFAC motifs, motif gene-set enrichment analysis, motif positional bias, 3′-UTR 8-mer discovery, identification of new miRNA genes, estimating number of miRNA targets, and experimental verification of miRNA genes.

Received 23 November 2004; accepted 9 February 2005; doi:10.1038/nature03441.

Published online 27 February 2005.

1. Gumucio, D. L. *et al.* Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human gamma and epsilon globin genes. *Mol. Cell. Biol.* **12**, 4919–4929 (1992).
2. Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W. & Lawrence, C. E. Human-mouse genome comparisons to locate regulatory sites. *Nature Genet.* **26**, 225–228 (2000).
3. Dubchak, I. *et al.* Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res.* **10**, 1304–1306 (2000).
4. Pennacchio, L. A. *et al.* An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science* **294**, 169–173 (2001).
5. Boffelli, D. *et al.* Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**, 1391–1394 (2003).
6. Sandelin, A., Wasserman, W. W. & Lenhard, B. ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res.* **32**, W249–W252 (2004).
7. Sinha, S., Blanchette, M. & Tompa, M. PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics* **5**, 170 (2004).

8. Dermitzakis, E. T. *et al.* Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. *Genome Res.* **14**, 852–859 (2004).
9. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).
10. Eddy, S. R. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.* **3**, e10 (2005).
11. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254 (2003).
12. Cliften, P. *et al.* Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301**, 71–76 (2003).
13. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
14. International Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
15. Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521 (2004).
16. Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004).
17. Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P. & Burge, C. B. Prediction of mammalian microRNA targets. *Cell* **115**, 787–798 (2003).
18. Maglott, D. R., Katz, K. S., Sicotte, H. & Pruitt, K. D. NCBI's LocusLink and RefSeq. *Nucleic Acids Res.* **28**, 126–128 (2000).
19. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14**, 708–715 (2004).
20. Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003).
21. Mootha, V. K. *et al.* Errα and Gabpa/b specify PGC-1α-dependent oxidative phosphorylation gene expression that is altered in diabetic muscle. *Proc. Natl Acad. Sci. USA* **101**, 6570–6575 (2004).
22. Johnston, S. D. *et al.* Estrogen-related receptor alpha 1 functionally binds as a monomer to extended half-site sequences including ones contained within estrogen-response elements. *Mol. Endocrinol.* **11**, 342–352 (1997).
23. Matys, V. *et al.* TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* **31**, 374–378 (2003).
24. Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA* **101**, 6062–6067 (2004).
25. Kuersten, S. & Goodwin, E. B. The power of the 3′ UTR: translational control and development. *Nature Rev. Genet.* **4**, 626–637 (2003).
26. Lai, E. C. Micro RNAs are complementary to 3′ UTR sequence motifs that mediate negative post-transcriptional regulation. *Nature Genet.* **30**, 363–364 (2002).
27. Lim, L. P., Glasner, M. E., Yekta, S., Burge, C. B. & Bartel, D. P. Vertebrate microRNA genes. *Science* **299**, 1540 (2003).
28. Griffiths-Jones, S. The microRNA Registry. *Nucleic Acids Res.* **32** (Database issue), D109–D111 (2004).
29. Fontana, W. *et al.* RNA folding and combinatorial landscapes. *Phys. Rev. E* **47**, 2083–2099 (1993).
30. Hofacker, I. L. Vienna RNA secondary structure server. *Nucleic Acids Res.* **31**, 3429–3431 (2003).
31. Ambros, V. *et al.* A uniform system for microRNA annotation. *RNA* **9**, 277–279 (2003).
32. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15–20 (2005).
33. Lim, L. P. *et al.* Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* advance online publication, 30 January (2005) (doi:10.1038/nature03315).
34. Berezikov, E. *et al.* Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* **120**, 21–24 (2005).
35. Chen, C. Y. & Shyu, A. B. AU-rich elements: characterization and importance in mRNA degradation. *Trends Biochem. Sci.* **20**, 465–470 (1995).
36. Spassov, D. S. & Jurecic, R. The PUF family of RNA-binding proteins: does evolutionarily conserved structure equal conserved function? *IUBMB Life* **55**, 359–366 (2003).
37. Margulies, E. H. *et al.* An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc. Natl Acad. Sci. USA* (in the press).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank B. Birren, M. Kamal, K. O'Neill and A. Subramanian for advice and discussions. This work was supported in part by grants from the National Human Genome Research Institute.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.K. (manoli@mit.edu) or E.S.L. (lander@broad.mit.edu).

Tropical to mid-latitude snow and ice accumulation, flow and glaciation on Mars

J. W. Head¹, G. Neukum², R. Jaumann³, H. Hiesinger¹, E. Hauber³, M. Carr⁴, P. Masson⁵, B. Foing⁶, H. Hoffmann³, M. Kreslavsky¹, S. Werner², S. Milkovich¹, S. van Gasselt² & The HRSC Co-Investigator Team*

¹Department of Geological Sciences, Brown University, Providence, Rhode Island 02912, USA

²Institut fuer Geologische Wissenschaften, Freie Universitaet Berlin, Malteserstrasse 74-100, Bldg D, 12249 Berlin, Germany

³DLR-Institut fuer Planetenforschung, Rutherfordstrasse 2, 12489, Berlin, Germany

⁴US Geological Survey, MS 975, Menlo Park, California 94025, USA

⁵Orsay-Terre, F-91405, Orsay Cedex, France

⁶ESTEC/SCI-SR, Postbus 299, NL2200AG Noordwijk, The Netherlands

*A list of all members of The HRSC Co-Investigator Team and their affiliations appears at the end of the paper

Images from the Mars Express HRSC (High-Resolution Stereo Camera) of debris aprons at the base of massifs in eastern Hellas reveal numerous concentrically ridged lobate and pitted features and related evidence of extremely ice-rich glacier-like viscous flow and sublimation. Together with new evidence for recent ice-rich rock glaciers at the base of the Olympus Mons scarp superposed on larger Late Amazonian debris-covered piedmont glaciers, we interpret these deposits as evidence for geologically recent and recurring glacial activity in tropical and mid-latitude regions of Mars during periods of increased spin-axis obliquity when polar ice was mobilized and redeposited in microenvironments at lower latitudes. The data indicate that abundant residual ice probably remains in these deposits and that these records of geologically recent climate changes are accessible to future automated and human surface exploration.

Among the most sensitive abiotic indicators of climate change are the accumulation, stability and flow of snow and ice. During the Little Ice Age on Earth (late sixteenth to early twentieth centuries), for example, glaciers at high latitude and altitude advanced an average of several kilometres¹ and today many are receding in concert with warming trends². On Mars, shallow subsurface water-ice stability in the current climate is limited to latitudes higher than about 60°, a theoretical prediction³ borne out by spacecraft observation⁴. At present the spin-axis obliquity of Mars, thought to be among the major factors in climate change, is about 25°, but calculations show that there were several periods of increasingly higher obliquity in the last several millions of years of the history of Mars⁵. General circulation models show that increased obliquity warms ice-rich polar regions and redistributes water-ice deposits equatorward^{6–8}. Indeed, geological observations show evidence for a recent ice age in the last several million years in the form of a latitude-dependent dust–ice mantle extending from high latitudes down to about 30° latitude in both hemispheres⁹, and evidence for localized tropical mountain glacier deposits that formed during earlier epochs of the Late Amazonian period on Mars tens to hundreds of millions of years ago¹⁰. Furthermore, there are numerous morphologic features that might involve ice-rich material at low to mid-latitudes throughout the history of Mars (such as landslides, debris aprons, rock glaciers and piedmont glaciers) but the origins, sources, amounts and state of water in these materials has been controversial^{11–20}.

Here we report on results from the High Resolution Stereo Camera (HRSC)²¹ on board Mars Express that show evidence for (1) the presence of significant volumes of ice and glacial-like flow in massif-marginal deposits at low to mid-latitudes (east of Hellas basin), and (2) very young glacier deposits in equatorial regions (Olympus Mons), suggesting recent climate change. Together these deposits are testimony to the importance and scale of equatorward water redistribution during recent climate change, and to the high

likelihood of the presence of significant volumes of buried ice currently in low-latitude regions on Mars.

Glacial-like flow in debris aprons

Debris aprons are a class of geomorphic features seen in mid-latitudes of Mars that are hundreds of metres thick, slope gently away from scarps or highland massifs, terminate at lobate margins, and are interpreted to be viscous flow features of material containing some portion of lubricating ice derived from adjacent highlands¹¹. New altimetry and high-resolution images have permitted more comprehensive observations and modelling but have not been able to distinguish conclusively among multiple models of apron formation¹⁷ (for example, ice-assisted rock creep, ice-rich landslides, rock glaciers and debris-covered glaciers) because of our inability to determine the proportion of ice in the rock debris (which can range widely, from ice deposited in debris interstices to debris deposited on ice accumulations)^{11–17}. Indeed, different aprons may have different modes of formation.

New HRSC data provide wide coverage of high-resolution data with a high signal-to-noise ratio and stereo capability. Analysis of new HRSC data of a massif-marginal lobe in the eastern Hellas region conclusively shows that the proportion of ice in this deposit was substantial enough to signify glacial and debris-covered glacial activity. Specifically, an 18-km-wide lobe extends about 8 km from the base of a 3.75-km-high massif (Fig. 1). The lobe is up to about 250 m thick, has a convex upward topographic profile, and is separated from the base of the massif by an irregular 50–100-m-deep depression (Fig. 1b, c). A broad alcove in the massif (Fig. 1a) adjacent to the lobe could be interpreted as a landslide scar, representing the source region for the lobate deposit. However, we find several inconsistencies with such an interpretation. For example, within the lobe itself (Fig. 1d), a distal 4-km zone is characterized throughout by a fretted and honeycomb-like texture of irregular pits and ridges. Depressions typically 20–40 m deep

make up 30–40% of this zone and occur between linear moraine-like ridges forming broad convex-outward lobes.

These patterns of sinuous ridges and irregular depressions are unlike landslides and are typical of Earth glacial deposits that remain following debris-containing glacial ice advance, stagnation and ablation²². Debris input to glaciers occurs most commonly at ice margins and is thus concentrated along the base of cirques and in medial debris septa that ultimately become medial moraines²³. Proximal debris addition from rockfalls and increasing debris concentration from below by ice sublimation results in supraglacial debris mantles in the distal direction with great spatial variability in thickness and grain size. As ablation proceeds, debris accumulations represented by englacial septa emerge and form longitudinal or transverse debris ridges separated by areas of cleaner or bare ice. Continued ablation results in the downwasting of the cleaner ice to produce pits, dirt cones and topographic inversion between the ridges, and ultimately the emerging or redistributed debris becomes thick enough to retard further sublimation²⁴.

The morphological similarities of features observed in Fig. 1 to glacial deposits are thus suggestive of processes of snow and ice accumulation, viscous flows of debris-containing ice, and the subsequent sublimation of significant volumes of the ice in the deposit, leaving behind numerous large sublimation pits tens of metres deep and intervening morainal ridges. Towards the massif, the presence of the linear depression suggests that this may have been the region of snow and ice accumulation; the present depressed topography (with local pits approaching a depth of 100 m) may represent typical proximal high concentrations of ice where more complete sublimation would take place. The presence of numerous plateaus in the fretted part of the deposit suggests that substantial quantities of ice remain beneath a local debris-rich cover, while sublimation has removed intervening ice to produce fretted pits.

These relationships strongly suggest that rather than a landslide scar, the broad alcove represents an accumulation zone for snow and ice that incorporates debris from the massif and under appropriate accumulation conditions, flows out into the surrounding terrain. Although we can see evidence for older lobate deposits, no superposed impact craters have been observed on this deposit, suggesting

a geologically recent age for this deposit. Despite its apparent recent age, the evidence for extensive sublimation and wasting are strong indicators that the current environment is not conducive to this large-scale accumulation and flow. However, the probable preservation of ice beneath a substantial portion of this deposit (underlying the sublimation till surface of the inter-pit plateaus) supports the probability that many of the other debris aprons also represent very ice-rich debris-covered glaciers. Indeed, theoretical predictions for Mars imply that dust can insulate buried ice²⁵ and observations in the Antarctic Dry Valleys suggest that sublimation tills can protect underlying glacial ice for millions of years²⁶.

Hour-glass-shaped flow in craters

Do other examples of debris aprons show features that might further distinguish between origins from ice-assisted rock creep, ice-rich landslides, rock glaciers and debris-covered glaciers?^{11–17} We see further evidence for viscous flow of very ice-rich material in the HRSC data of an hourglass-shaped deposit occurring in two craters at the base of a 3.5–4-km-high massif located on the eastern rim of the Hellas basin (Fig. 2). Two adjacent circular depressions about 9 and 16 km in diameter extend outward from the base of the massif into the surrounding lowlands. In contrast to the previously described deposit, which was oriented with its long axis parallel to the massif slope (Fig. 1), in this location (Fig. 2) the deposit is contained within the craters and appears to fill them. In the proximal crater, the floor is a regionally flat surface that lies nearly at the rim, about 500 m above the surrounding plain, and along a N–S topographic profile (Fig. 2e) the crater appears to be filled nearly to the brim. E–W profiles show, however, that the floor tilts away from the massif at a slope of less than a few degrees (Fig. 2f).

The surface texture on the floor of the crater revealed by the HRSC data (Fig. 2c, d) shows unequivocal evidence for streamlines and lobes typical of ice flow and ice-lobe interaction. Four discrete zones are seen within this crater. To the north, arcuate nested lobes emerge fully developed from the base of the slope and are progressively compressed along the margin; these give way in the north-central part of the crater to a 2.5-km-wide zone of parallel ridges that converge to a narrower 1-km-wide zone where the rim of the

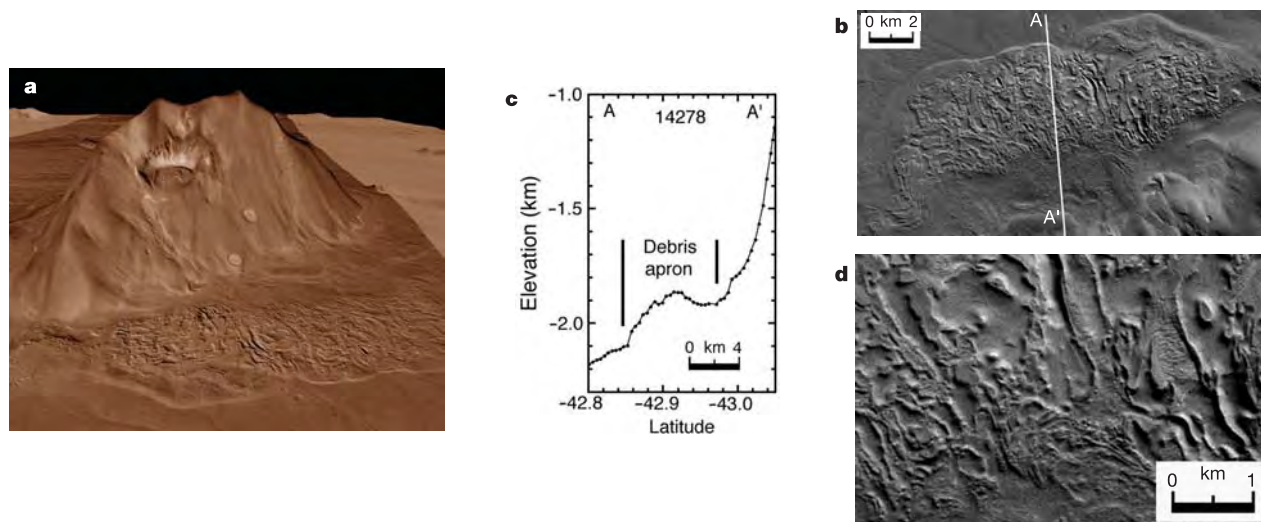


Figure 1 Massif and debris-apron deposits on the eastern rim of the Hellas basin (262.8° W, –43.2°). All images are portions of MEX HRSC Orbit 300. **a**, Perspective view looking east of the 3.75-km-high massif surrounded by debris aprons. Vertical exaggeration is ~30×. **b**, Debris apron with a significant portion comprised of fretted pits and depressions, suggesting the former presence of ice. Line shows location of MOLA

profile A–A'. **c**, Altimetric profile across base of massif and debris-apron deposit. MOLA orbit 14278. **d**, Enlargement of portion of fretted debris-apron deposit located just to the right of the profile (Fig. 1b). Note the lobate shapes and the presence of lateral and arcuate ridges and central depressions.

proximal crater is breached. A south central 1.5-km-wide zone of tight arcuate lobes points downslope and becomes compressed into parallel flowlines as it reaches the distal breach. In the southern part of the crater, a 3-km-wide zone of parallel ridges is progressively compressed into a very narrow zone less than 1 km in width. All four of these zones join together at the low point in the crater rim, and flow through a narrow breach less than 2 km wide (Fig. 2d), dropping several hundred metres in elevation and spreading out onto the lower crater floor, creating a set of lobate ridges and depressions further indicative of viscous flow. This configuration is very similar to Earth glacial environments, such as the 60-km-wide Malispina glacier²⁷, where parallel ridges form due to low-viscosity,

high-strain-rate flow in narrow valley glaciers; these then emerge out onto a broad plain and spread out to form a piedmont glacier many times the width of the initial constriction.

The viscous-flow-like crater-filling materials appear to be fully developed at the proximal end of the smaller crater adjacent to the massif alcove (Fig. 2a–d). Here too, the distinctive alcove in the massif could, in principle, represent a landslide scar. Examination of the alcove area details, however, reveals evidence for individual topographic lineations and depressions favouring the accumulation and flow of ice, rather than landslides²⁰. In summary, the nature, morphology and topography of the deposit indicates that the alcove served as an accumulation zone for snow and ice that acquired a debris cover from the surrounding steep slopes, and flowed out from the base of the alcove into the surrounding depressions, filling the proximal one and then breaching and overflowing to fill the lower depression. Further evidence that the viscously flowing material was predominantly ice comes from (1) the expanded lobate and complexly deformed nature of the deposit as it spreads out onto the floor of the lower crater from the notch in the upper crater (Fig. 2a, b), (2) the abundance of irregular-shaped pits in the distal lobes, indicating ice sublimation (Fig. 2a, b), (3) evidence for distal moraines around the crater interior margins, indicating ice retreat (Fig. 2b), and (4) compressed and deformed ridges and elongated craters (Mars Orbiter Camera (MOC) image M2300829) also suggest an ice-like rheology (Fig. 2c, d). We thus interpret these features to be debris-covered piedmont-type glaciers.

Young rock glaciers at Olympus Mons

Evidence for extensive debris-covered piedmont glaciers along the northwest edge of the Olympus Mons scarp has been described from Viking and THEMIS data^{20,28,29}, and together with the tropical mountain glaciers at the Tharsis Montes^{10,30}, they stand as evidence of extensive localized glaciation in the Late Amazonian when obliquity was typically⁵ in excess of 35°. These glaciers extend up to 70–120 km from the base of the Olympus Mons scarp^{28,29} and over 350 km from Arsia Mons¹⁰.

HRSC data reveal the presence of several additional rock-glacier-like features along the Olympus Mons basal scarp that are clearly superposed on top of the larger, and thus older, debris-covered piedmont glaciers (Fig. 3). Here the individual lobes are about 25 km in length, and much narrower than the piedmont glacier

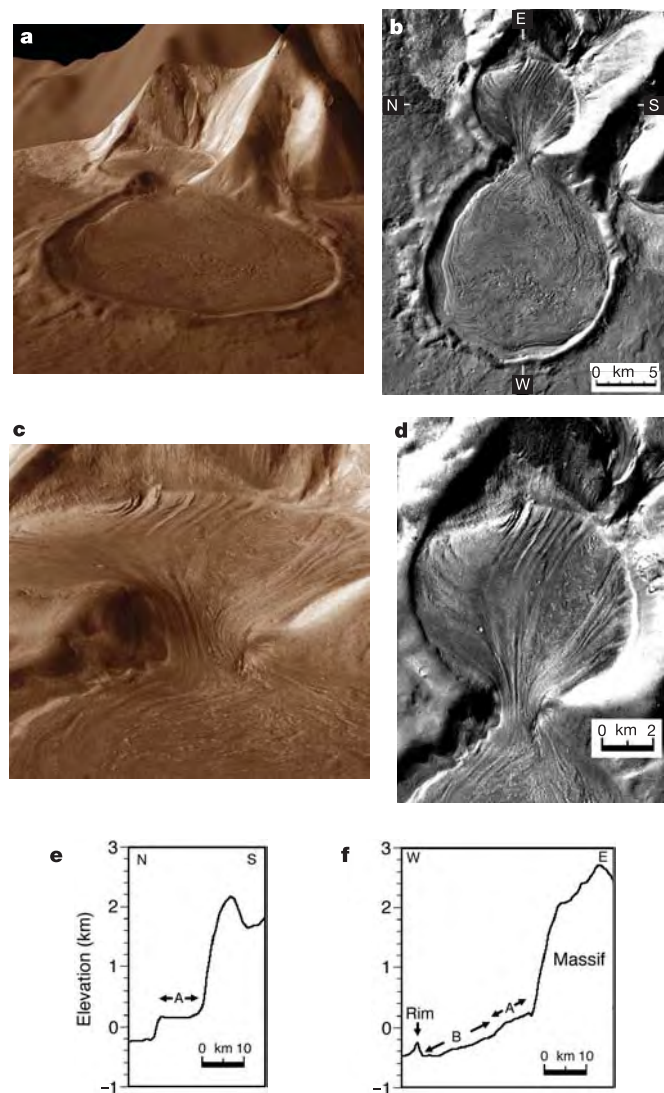


Figure 2 Hourglass-shaped deposit at the base of a massif on the eastern Hellas basin rim (257° W, –39.2°). All images are from MEX HRSC orbit 248. **a**, Perspective view of the 3.5–4-km-high massif showing viscous flow of material from a 9-km crater (marked A in Fig. 2e, f) through a narrow notch into a 16-km-diameter crater (marked B in Fig. 2f). Vertical exaggeration is ~30×. **b**, Vertical view of the two craters showing flowlines and lobes. Tickmarks show the location of MOLA gridded topography profiles in **e** and **f**. **c**, Enlargement of the notch between the two craters showing the four zones where flowlines starting at the base of the slope converge at the narrow 2-km-wide notch and then diverge and spread laterally out into the lobate deposits below. **d**, Perspective view of the proximal crater showing flowlines converging in the gap. MOLA gridded altimetric profiles N–S (**e**) and E–W (**f**).

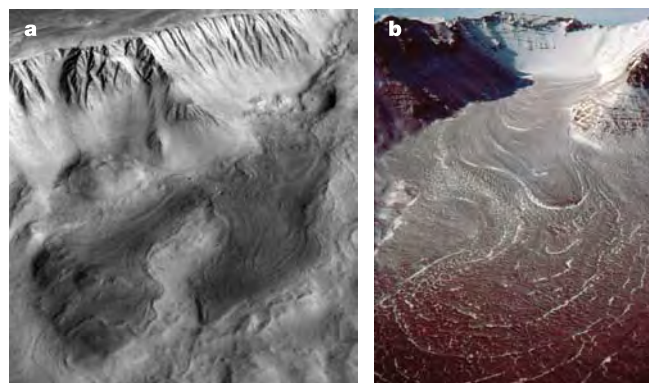


Figure 3 Deposits from a recent lobate rock glacier at the base of the Olympus Mons scarp (138° W, 18°). **a**, Perspective view looking southwest towards the 6-km-high scarp. Note lobate deposits extending about 20–25 km from the base of the alcoves (from right centre towards lower left) darkened for emphasis. HRSC data from MEX orbit 143. **b**, Perspective view of the upper ~5 m of a debris-covered rock glacier emerging from a cirque in Mullins Valley, Antarctic Dry Valleys on Earth. Note the morphologic similarity to features at the base of Olympus Mons. Photo courtesy of David Marchant.

deposit on which they are superposed (Fig. 3a). Furthermore, the paths of the lobes closely follow the topography of the pre-existing deposit, an indication that they are the result of material advancing from the base outward, and not just backwasting and retreat of the residual larger lobe. The sources of the lobes are cusate alcoves in the basal scarp (Fig. 3a), topographic depressions that are natural traps for wind-blown snow in terrestrial rock glacier environments^{31–33} (Fig. 3b). Previously, the broad lobate features had been interpreted to be landslides³⁴, but terrestrial analogues²⁰ and new data^{10,28–30} have provided very strong evidence for a debris-covered piedmont glacier origin. The new HRSC data add further support to the general glacial interpretation; evidence supporting a rock glacier interpretation for these smaller lobate features includes (1) their origin in alcoves, (2) their elongated subparallel concentric ridges, distorted in relation to underlying and adjacent topography, (3) distinctive terminal moraines, and (4) their strong morphological similarities to terrestrial rock glaciers of known glacial origin^{10,30–32} (compare Fig. 3a, b).

Thus, we conclude that there is clear evidence for the formation

of recent rock glaciers locally at the base of the Olympus Mons scarp (Fig. 3), similar to typical rock glaciers formed in cirques in the Antarctic Dry Valleys, and their recent advance on Mars to distances measured in several tens of kilometres. The lack of snow in the present alcove, evidence of a depression there, and the presence of fans of scarp talus spreading downslope into these regions with no evidence of gelifluction (the slow downslope movement of sediment associated with seasonal thawing of ground ice), are all evidence that the young rock glaciers are no longer active and that the snow-accumulation conditions that led to their formation no longer persist in this area of Mars.

Role of glaciation in debris-apron origin

Previously, significant controversy surrounded the origin of debris aprons at the base of many massifs on Mars. Outstanding questions focused on (1) the abundance of ice in the deposits during formation, (2) the origin of this ice ('bottom up', from ground ice or groundwater, or 'top down', from atmospheric frost or snow accumulation), and (3) the mode of origin of these features (ranging

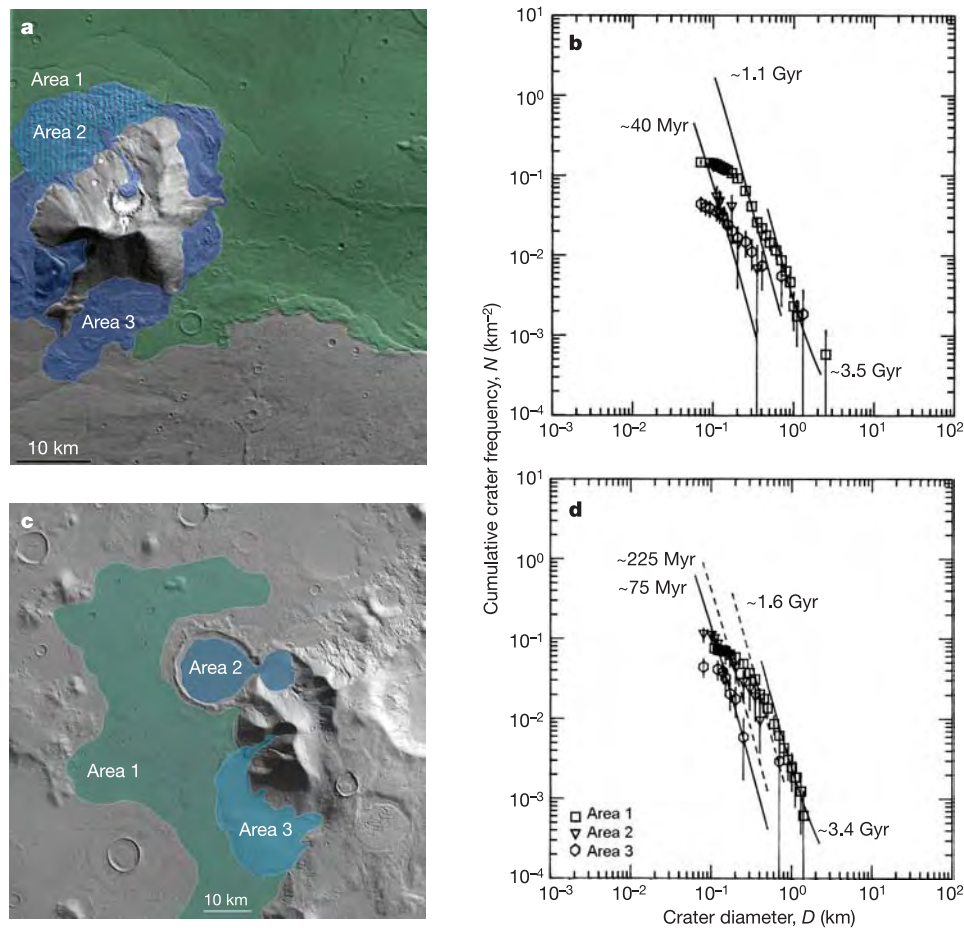


Figure 4 Ages of events in the lobate debris aprons. The impact-crater size-frequency distribution^{37–39} has been fitted to individual segments of the distribution, giving values for the different episodes; application of the Hartmann–Neukum cratering chronology provides absolute ages. For ages younger than ~3 Gyr the error is around 20–30%. Error bars are one-sigma. All ages <2 Gyr may be affected in the same way by a possible systematic error of ~2× in the cratering model used³⁹. **a, b**, Massif and fretted debris apron (**a** and Fig. 1), and impact crater size-frequency distributions (**b**). Area 1, surrounding the massif and lobate deposit, has an age of 3.5 Gyr and appears to have been resurfaced as late as 1.1 Gyr ago, due in part to fluvial activity represented by numerous channels. Debris aprons surrounding the massif (area 3 combined with area 2) display evidence for almost continuous erosion. The portion of the debris apron deposit

represented by the fretted debris apron (area 2) is as young as 40 Myr. If debris apron fretting is due to ice sublimation and the sublimation process was continuous over millions of years, some impact craters might be lost and therefore 40 Myr would be a minimum age. Crater distributions are from HRSC orbit 300 data. **c, d**, Hourglass structure apron (**c** and Fig. 2), and impact crater size-frequency distributions (**d**). The area surrounding the hourglass and lobate deposit (area 1) has an age of 3.4 Gyr and appears to have been almost continuously resurfaced, with a pause at ~1.4 Gyr, and with some indication of a resurfacing event ending about 225 Myr ago (second dashed line). The hourglass deposit (area 2) and the adjacent debris apron to the south (area 3) have a well-defined age of 75 Myr. Crater distributions from HRSC orbit 248 data.

from ice-assisted rock creep, ice-rich landslides, rock glaciers, to debris-covered glaciers^{11–17}. HRSC data provide important new evidence relevant to each of these outstanding questions. First, the HRSC data support a dominant role for ice in the formation of some of these debris aprons and related deposits with evidence from the pitted terrain indicating very high proportions of ice in a debris apron (Fig. 1b, d), and evidence for low-viscosity flow in the hourglass-like feature, indicating high ice–debris ratios (Fig. 2). Second, evidence for the very high abundance of ice in the deposits and hyperarid cold polar-desert-like conditions during formation argues against bottom-up sources such as groundwater, and favours top-down sources, such as atmospheric-related processes of frost, snow or ice accumulation. Third, the proximity of the source regions for these features to steep-sided debris-covered alcoves, and their close analogue to areas on Earth where snow accumulates to form debris-covered, viscously flowing ice deposits (glaciers) (Fig. 3), strongly suggests that many of these features originate as extremely ice-rich, debris-covered glaciers. The geometry of the features discovered along the base of the Olympus Mons scarp, their similar origin in alcoves, their similarities in size and morphology to terrestrial debris-covered ice-rich rock glaciers, and their direct superposition on larger deposits interpreted to be Late Amazonian debris-covered piedmont glaciers^{28,29} all strengthen the interpretation of the potential importance of glacial activity in the formation and evolution of debris aprons surrounding massifs on Mars.

Conclusions and implications

The presence of significant volumes of ice and glacial-like flow in massif-marginal deposits at low to mid-latitudes on Mars (east of Hellas basin; -39° to -43° latitude) and in rock glaciers at the base of Olympus Mons ($+18^{\circ}$) strongly suggests that conditions in the geological past have favoured the accumulation of snow and ice and its flow in these tropical and mid-latitude regions. Crater size-frequency distribution data collected from the HRSC images (Fig. 4) of the lobate debris aprons east of Hellas (Figs 1, 2) show evidence for multiple eras of ice-related resurfacing, while the Olympus Mons rock glaciers (Fig. 3) are just a few million years old³⁵. Thus, these deposits are further testimony to the importance and scale of equatorward water redistribution during climate change^{6–8}, and its accumulation in specific areas^{35,36}. Furthermore, the superposition of the Olympus Mons rock glaciers on older debris-covered piedmont glacier deposits^{28–29}, dated at 280 to 130 million years (Myr) ago for the major lobes³⁶, strongly suggests that these conditions fluctuate with time, and that the geologically very young Olympus Mons rock glaciers documented here (Fig. 3a) represent a recent return to these conditions a few million years ago³⁶ for periods shorter than those that formed the underlying, much more extensive, Late Amazonian deposits. Finally, that none of these features at present seem to be accumulating ice, and thus flowing and advancing, but instead appear to be undergoing sublimation and wasting, strongly suggests that conditions conducive to their formation are not currently in effect. The latter point is consistent with the idea that Mars may now be in an ‘interglacial’ period due to its relatively low obliquity⁹. Thus, deposits such as these revealed in detail by the HRSC data provide an important geological record of recent climate change that can be used to test and improve both models of recent climate change^{6–8} and predictions of the history of orbital parameters⁵. The lower latitudes of these ice-rich deposits also mean that this key climate record is very accessible to automated and human exploration for direct examination and analysis. □

Received 4 October 2004; accepted 13 January 2005; doi:10.1038/nature03359.

1. Grove, J. M. *The Little Ice Age* (Routledge, London, 1988).
2. Warren, C. R. Glaciers in the greenhouse. *Geogr. Rev.* **8**, 2–7 (1995).

3. Mellon, M. T. & Jakosky, B. M. The distribution and behavior of Martian ground ice during past and present epochs. *J. Geophys. Res.* **100**, 11781–11799 (1995).
4. Feldman, W. C. *et al.* Global distribution of neutrons from Mars: Results from Mars Odyssey. *Science* **297**, 75–78 (2002).
5. Laskar, J. *et al.* Long term evolution and chaotic diffusion of the insolation quantities of Mars. *Icarus* **170**, 343–364 (2004).
6. Richardson, M. I. & Wilson, R. J. Investigation of the nature and stability of the Martian seasonal water cycle with a general circulation model. *J. Geophys. Res.* **107**, doi:10.1029/2001JE001536 (2002).
7. Mischna, M. *et al.* On the orbital forcing of Martian water and CO₂ cycles: A general circulation model study with simplified volatile schemes. *J. Geophys. Res.* **108**, doi:10.1029/2003JE002051 (2003).
8. Haberle, R. M. *et al.* Orbital change experiments with a Mars general circulation model. *Icarus* **161**, 66–89 (2003).
9. Head, J. W. *et al.* Recent ice ages on Mars. *Nature* **426**, 797–802 (2003).
10. Head, J. W. & Marchant, D. R. Cold-based mountain glaciers on Mars: Western Arsia Mons. *Geology* **31**, 641–644 (2003).
11. Squires, S. W. Martian fretted terrain—Flow of erosional debris. *Icarus* **34**, 600–613 (1978).
12. Colaprete, A. & Jakosky, B. M. Ice flow and rock glaciers on Mars. *J. Geophys. Res.* **103**, 5897–5909 (1998).
13. Mangold, N. & Allemand, P. Topographic analysis of features related to ice on Mars. *Geophys. Res. Lett.* **28**, 407–410 (2001).
14. Mangold, N. *et al.* Experimental and theoretical deformation of ice-rock mixtures: Implications on rheology and ice content of Martian permafrost. *Planet. Space Sci.* **50**, 385–401 (2002).
15. Baratoux, D. *et al.* Evidence of liquid water in recent debris avalanche on Mars. *Geophys. Res. Lett.* **29**, doi:10.1029/2001GL014155 (2002).
16. Mangold, N. Geomorphic analysis of lobate debris aprons on Mars at Mars Orbiter Camera scale: Evidence for ice sublimation initiated by fractures. *J. Geophys. Res.* **108**, doi:10.1029/2002JE001885 (2003).
17. Pierce, T. L. & Crown, D. A. Morphologic and topographic analyses of debris aprons in the eastern Hellas region, Mars. *Icarus* **163**, 46–65 (2003).
18. Kargel, J. S. & Strom, R. G. Ancient glaciation on Mars. *Geology* **20**, 3–7 (1992).
19. Baker, V. R. *et al.* Ancient oceans, ice sheets and the hydrological cycle on Mars. *Nature* **352**, 589–594 (1986).
20. Lucchitta, B. K. Mars and Earth—Comparison of cold-climate features. *Icarus* **45**, 264–303 (1981).
21. Neukum, G. *et al.* HRSC: The High Resolution Stereo Camera of Mars Express 17–35 (Report ESA-SP-1240, European Space Agency Publications Division, Noordwijk, The Netherlands, 2004).
22. Benn, D. I. & Evans, D. J. A. *Glaciers and Glaciation* 237–239 (Arnold, London, 1998).
23. Eyles, N. & Rogerson, R. J. A framework for the investigation of medial moraine formation: Austerdalsbreen, Norway, and Berendon Glacier, British Columbia, Canada. *J. Glaciol.* **20**, 99–113 (1978).
24. Boulton, G. S. in *Till: A Symposium* (ed. Goldthwait, R. P.) 41–72 (Ohio State Univ. Press, Columbus, 1971).
25. Skorov, Yu. V. *et al.* Stability of water ice under a porous nonvolatile layer: implications to the south pole layered deposits of Mars. *Planet. Space Sci.* **49**, 59–63 (2001).
26. Marchant, D. *et al.* Formation of patterned ground and sublimation till over Miocene glacier ice in Beacon Valley, southern Victoria Land, Antarctica. *Geol. Soc. Am. Bull.* **114**, 718–730 (2002).
27. Hartshorn, J. H. Superglacial and proglacial geology of the Malaspina Glacier, Alaska, and its bearing on glacial features of New England. *Geol. Soc. Am. Bull.* **63**, 1259–1260 (1952).
28. Milkovich, S. M. & Head, J. W. Olympus Mons fan shaped deposit morphology: Evidence for debris glaciers. *6th Int. Mars Conf.* abstr. 3149 (2003).
29. Head, J. W., Shean, D. E., Milkovitch, S. M. & Marchant, D. Tropical mountain glaciers on Mars: Evidence for Amazonian climate change. *3rd Mars Polar Conf.* abstr. 8105 (2003).
30. Shean, D. E. *et al.* Tharsis Montes cold-based glaciers: Observations and constraints for modeling and preliminary results. *Lunar Planet. Sci.* **XXXV**, abstr. 1438 (2004).
31. Potter, N. Periglacial geomorphology. *J. Geol. Educ.* **32**, 226–232 (1984).
32. Johnson, P. G. Glacier-rock glacier transition in the southwest Yukon Territory. *Arctic Alpine Res.* **12**, 195–204 (1980).
33. Martin, H. E. & Whalley, W. B. Rock glaciers, part 1, Rock glacier morphology: Classification and distribution. *Prog. Phys. Geogr.* **11**, 260–282 (1987).
34. Morris, E. C. & Tanaka, K. L. *Geologic Maps of the Olympus Mons Region of Mars* (Map I-2327, Misc. Inv. Ser., US Geological Survey, Reston, Virginia, 1994).
35. Neukum, G. *et al.* Recent and episodic volcanic and glacial activity on Mars revealed by the High Resolution Stereo Camera. *Nature* **432**, 971–979 (2004).
36. Hauber, E. *et al.* Discovery of a flank caldera and very young glacial activity at Hecates Tholus, Mars. *Nature* doi:10.1038/nature03423 (this issue).
37. Neukum, G. *et al.* Cratering record in the inner Solar System in relation to the lunar reference system. *Space Sci. Rev.* **96**, 55–86 (2001).
38. Ivanov, B. A. Mars/Moon cratering rate ratio estimates. *Space Sci. Rev.* **96**, 87–104 (2001).
39. Hartmann, W. K. & Neukum, G. Cratering chronology and the evolution of Mars. *Space Sci. Rev.* **96**, 165–194 (2001).

Acknowledgements We thank S. Pratt, A. Cote and J. Dickson for help in data analysis and manuscript preparation, T. Roatsch for data handling, calibration and commanding, F. Scholten and K. Gwinner for photogrammetric processing, and V. Baker for a review. We thank the European Space Agency, DLR (German Aerospace Center), and the Freie Universitaet, Berlin, for their efforts in building and flying the HSRC experiment, and processing the data, and NASA for supporting the participation of J.W.H.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.W.H. (james_head@brown.edu).

The HRSC Co-Investigator Team: J. Alibert¹, A. T. Basilevsky², G. Bellucci³, J.-P. Bibring⁴, M. Buchroithner⁵, M. H. Carr⁶, E. Dorrer⁷, T. C. Duxbury⁸, H. Ebner⁹, B. H. Foing¹⁰, R. Greeley¹¹, E. Hauber¹², J. W. HeadIII¹³, C. Heipke¹⁴, H. Hoffman¹², A. Inada^{15,18}, W.-H. Ip¹⁶, B. A. Ivanov¹⁷, R. Jaumann¹², H. U. Keller¹⁸, R. Kirk¹⁹, K. Kraus²⁰, P. Kronberg²¹, R. Kuzmin², Y. Langevin⁴, K. Lumme²², W. Markiewicz¹⁸, P. Masson²³, H. Mayer⁷, T. B. McCord²⁴, J.-P. Muller²⁵, J. B. Murray²⁶, F. M. Neubauer²⁷, G. Neukum (Principal Investigator)²⁸, J. Oberst¹², G. G. Ori²⁹, M. Pätzold²⁷, P. Pinet³⁰, R. Pischel¹², F. Poulet⁴, J. Raitala³¹, G. Schwarz³², T. Spohn¹² & S. W. Squyres³³

Affiliations of authors: 1, TU Berlin, D-10623 Germany; 2, Vernadsky Institute-RAS, Moscow 117975, Russia; 3, CNR/IFSI, 00133 Rome, Italy; 4, CNRS/IAS, 91405 Orsay Cedex, France; 5, TU Dresden, D-01062 Germany; 6, USGS, MS 975 Menlo Park, California 94025, USA; 7, Universität der Bundeswehr München, D-85577 Neubiberg, Germany; 8, JPL, Pasadena, California 91109, USA; 9, TU München, D-80333 Germany; 10, ESTEC/SCI-SR, Postbus 299, NL-2200 AG Noordwijk, The Netherlands; 11, Arizona State University, Box 871404, Tempe, Arizona 85287-1404, USA; 12, DLR, D-12489 Berlin, Germany; 13, Brown University, Box 1846, Providence, Rhode Island 02912, USA; 14, Universität Hannover, D-30167 Germany; 15, Kobe University, 657-8501 Kobe, Japan; 16, National Central University (NCU), Chung-Le 320 Taiwan; 17, IDG-RAS, Moscow 117979, Russia; 18, MPAE, Postfach 20, D-37191 Katlenberg-Lindau, Germany; 19, USGS, Flagstaff, Arizona 86001; 20, TU Wien, A-1040 Wien, Austria; 21, TU Clausthal, D-38678 Clausthal, Germany; 22, University of Helsinki, PO Box 14, SF-00014 Helsinki, Finland; 23, Orsay Terre, F-91405 Orsay Cedex, France; 24, PSI-Nw, Winthrop, Washington 98862, USA; 25, University College London WC1E 6BT, UK; 26, The Open University, Milton Keynes MK7 6AA Buckinghamshire, UK; 27, Universität Köln, D-50923 Germany; 28, FU Berlin, D-12249, Germany; 29, IRSPS, I-65127 Pescara, Italy; 30, Observatoire de Midi-Pyrénées, F-31400 Toulouse, France; 31, University of Oulu, FIN-90401 Finland; 32, DLR, D-82234 Wessling, Germany; 33, Cornell University, Ithaca, New York 14853-1301, USA

Evidence from the Mars Express High Resolution Stereo Camera for a frozen sea close to Mars' equator

John B. Murray¹, Jan-Peter Muller², Gerhard Neukum³,
Stephanie C. Werner³, Stephan van Gasselt³, Ernst Hauber⁴,
Wojciech J. Markiewicz⁵, James W. Head III⁶, Bernard H. Foing⁷,
David Page^{1,8}, Karl L. Mitchell⁹, Ganna Portyankina⁵
& The HRSC Co-Investigator Team*

¹Department of Earth Sciences, The Open University, Milton Keynes MK7 6AA, UK

²Department of Geomatic Engineering, University College London, Gower Street, London WC1E 6BT, UK

³Geosciences Institute, Freie Universität Berlin, Malteserstrasse 74-100, Building D, 12249 Berlin, Germany

⁴DLR-Institut für Planetenforschung, Rutherfordstrasse 2, D-12489 Berlin-Adlershof, Germany

⁵Max Planck Institute for Aeronomy, Max-Planck-Strasse 2, 37191 Katlenburg-Lindau, Germany

⁶Department of Geological Sciences, Brown University, Box 1846, Providence, Rhode Island 02912, USA

⁷ESA Research and Scientific Support Department, ESTEC/SCI-SR postbus 299, 2200 AG Noordwijk, The Netherlands

⁸Department of Mineralogy, The Natural History Museum, London SW7 5PB, UK

⁹Environmental Science Department, Lancaster University, Bailrigg, Lancaster LA1 4YQ, UK

*A list of all members of The HRSC Co-Investigator Team and their affiliations appears at the end of the paper

It is thought that the Cerberus Fossae fissures on Mars were the source of both lava and water floods^{1,2-4} two to ten million years ago^{1,2,5}. Evidence for the resulting lava plains has been identified in eastern Elysium^{1,2,4,6-8}, but seas and lakes from these fissures and previous water flooding events were presumed to have evaporated and sublimed away⁹⁻¹¹. Here we present High Resolution Stereo Camera images from the European Space Agency Mars Express spacecraft that indicate that such lakes may still exist. We infer that the evidence is consistent with a frozen body of water, with surface pack-ice, around 5° north latitude and 150° east longitude in southern Elysium. The frozen lake measures about 800 × 900 km in lateral extent and may be up to 45 metres deep—similar in size and depth to the North Sea. From crater counts, we determined its age to be 5 ± 2 million years old. If our interpretation is confirmed, this is a place that might preserve evidence of primitive life, if it has ever developed on Mars.

Extensive fields of large fractured plate-like features on a horizontal surface are visible near the south end of the High Resolution Stereo Camera (HRSC) imaging strip taken on 19 January 2004 (Fig. 1). This area has previously been covered by NASA high-resolution Mars Orbiter Camera (MOC) imagery at pixel sizes down to 1.8 m (Fig. 2). The latter images show fractured plates at a smaller scale. Individual plates are of all sizes from 30 m up to >30 km, with clear signs of break-up, rotation (Fig. 1c) and horizontal drift for distances of several kilometres. The plates show characteristic differences from plate-like features elsewhere on Mars and in the east of Elysium Planitia. The latter features have been interpreted to be rafts of solidified lava floating on the surface of large flood basalts⁶, but several observations indicate that this cannot be the case in this area.

Surface ages were determined^{12,13} from the size-frequency distribution of 66 impact craters on HRSC images, which suggest a resurfacing event about 5 million years (Myr) ago. Counts of 268 craters on MOC images show that the plates are older than the brighter inter-plate areas (Fig. 3). The statistical errors of the two data sets (counts on plates and inter-plate areas) indicate that they

are almost coincident in age, but the whole inter-plate size-frequency distribution falls consistently below that for the plates. This is an indication that the inter-plate areas are really younger than the plates, but within the error limits the age difference could be from a few hundred thousand years up to 2 Myr, with a most likely value of 1 Myr. This age difference is independent of any systematic error in the cratering chronology model used^{12,13}. Basalt lava flows 50 m deep can remain partially molten at the centre for only about 5 yr (ref. 14), so these plates cannot be the result of surges and break-outs of lava carrying previously solidified crust, as occurred over timescales of a week or so during the 1783–84 Laki Fissure eruption, Iceland, which is the closest terrestrial analogue to martian flood lavas⁶.

Lava break-outs entail build-up and failure of inflating lava to create the plate-like morphology⁶, but there are no signs of the inflation that occurs on terrestrial basalts and in other areas of Mars. The Mars Orbiting Laser Altimeter (MOLA) topographic profiles across the area show a remarkably flat surface with broad topography varying by <5 m over more than 60 km, that is, a slope of <0.005°. This compares to a slope of about 0.2° for terrestrial flood basalts.

Furthermore, a drop in surface level occurred after flooding of 18 to 85 m (equivalent to about 9% to 16% of the depth before flooding) within flooded impact craters (Fig. 4). If this had been lava, such a drop would be impossible in these ponded enclosures, because thermal contraction of ponded lava would amount to less than 1% (ref. 15).

Other features in the HRSC image show unique features that provide a clue to their origin. Where the plates have drifted into obstacles, straight or curved lanes have formed downstream within the plates themselves ('L' and 'I' in Fig. 1c). These are not found within lava rafts. Also, the plates are one to two orders of magnitude larger than the largest-known terrestrial basalt rafts. Both these observations, together with the horizontal surface (<0.005°, corresponding to terrestrial tidal sea surface slopes in some estuarine situations) imply an extremely mobile fluid, with characteristics similar to those of water.

Other observations indicate the strong resemblance of these plates to pack-ice. Where pre-existing small topographic highs protrude through the plates to form islands, plate drift has caused rubble piles with pressure ridges on the upstream side (Figs 1c and 2a). Where the highs are craters, these ridges show a superficial resemblance to fluidized ejecta, but unlike the latter, the ridges are subconcentric to the rim and form on one side of the crater only (always the upstream side), show no lobate overlaps or signs of radial movement, have up to 20 subparallel ridges instead of one to three, and show no broad smoother areas proximal to the crater that indicate ejecta flow. Figure 2a shows pressure ridges (denoted 'R') within the rubble pile on the right with wavelengths between 10 and 70 m, which appear to have extended outward from the crater edge as the liquid level dropped and the frozen surface was grounded progressively further down the outer slopes of the crater. These are strikingly similar to rubble piles of sea ice that form around islands in the Arctic and Antarctic (Fig. 2b). The sagging and consequent surface cracking 'C' within the crater itself as the level dropped are also visible. One plate 'F' has drifted into the crater when the level was higher through the gap in the rim 'G', but then become grounded in its present position as the surface lowered, draping it over the northeast rim.

These craters and islands have acted in a similar manner to ice-breakers as the plates drifted past them, leaving straight or curved leads downstream with uniform width (Fig. 1c). The high-resolution MOC image in Fig. 2a shows that these lanes are still very smooth at the 10-m scale, as are similar features in pack-ice on Earth. In places the plates have moved in channels between zones of more stable ice, and overall the direction of drift is towards the west or southwest.

Combinations of processes are unlikely, and have no terrestrial counterparts. Lava plates rafted on mud are not possible because basalt has a density >70% greater than mud. Lava on ice would create pseudocraters, melting and consequent sagging of the raft centres, none of which is observed in this area, and mud rafts floating on mudflows would sink, solid mud having a density 10% greater than fluid mud.

On the basis of this analysis, we interpret the structures and textures to be due to pack-ice formed as a moving and fracturing thermal boundary layer on top of ponded aqueous floodwater that later froze. An early drop in water level occurred while the ice was still drifting (Fig. 2a), mainly owing to evaporation/sublimation, or perhaps seepage of liquid water into the substratum.

Reasonable estimates of the depth can be made by using the rim height to diameter ratios of submerged impact craters. Mean values for simple martian craters are given by: $H = 0.04D^{0.31}$, where H is exterior rim height and D is crater diameter¹⁶. The impact crater 'I' (Fig. 1c) is 1.1 km in diameter and has only the highest part of the

rim exposed, suggesting an initial water depth of about 42 m at that point. Fourteen other crater rims have been identified from their traces partially above or just below the ice as the surface has lowered, yielding initial water depths of between 31 and 53 m, with an average at 45 m. The true depth may be less than these values, as some suspended sediment transported in the early stages would have settled out before freezing. The MOLA profiles across three flooded craters indicate that low parts of the rim are still 0 to 30 m above the mean ice level, suggesting that the evaporation, sublimation and seepage sagging referred to above may have lowered the ice thickness to a present mean depth of around 30 m.

The area lies at the foot of the Athabasca Valles system, where sinuous ridges within the complex of valleys have been likened to those of pack-ice, and plate-like structures there have been proposed to be casts of sediment-rich ice deposited by an ice-rich debris flow, the ice having later sublimated away¹⁷. Fluvial bedforms also indicate that the Athabasca Valles contained water channels, perhaps fed by a large over-pressured subsurface aquifer¹⁸ giving rise to high

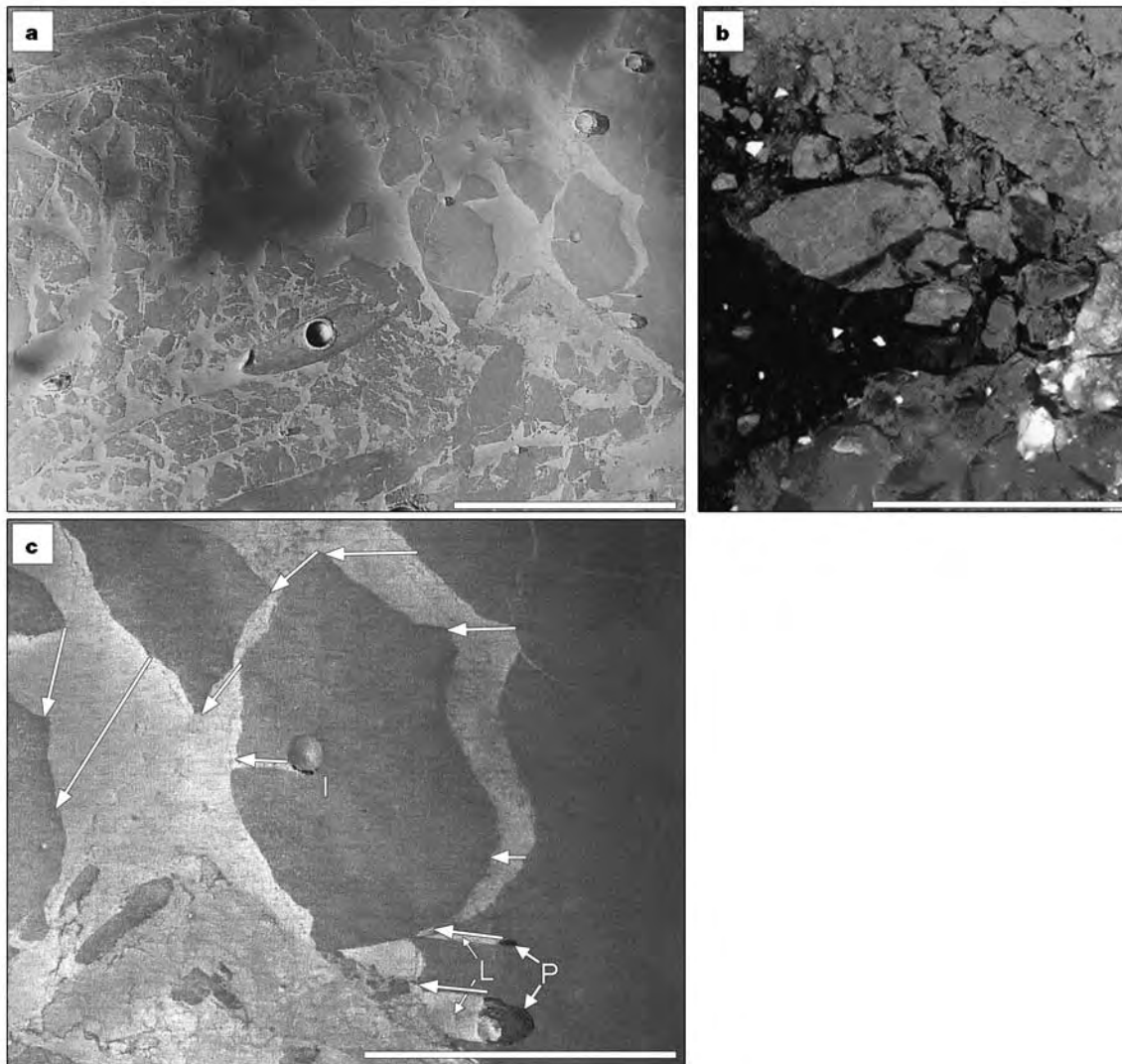


Figure 1 Views of plate-like terrain on Mars, and pack-ice on Earth. **a**, Part of an HRSC image of Mars from orbit 32, with a resolution of 13.7 m per pixel, centred at 5.5° north latitude and 150.4° east longitude, showing plate-like deposits with signs of break-up, rotation and lateral movement to the west-southwest in the lower part of the image. Scale bar is 25 km. **b**, Synthetic Aperture Radar image of pack-ice in the Weddell Sea, Antarctica. Scale bar is 25 km. (ESA image, processed by H. Rott.) **c**, Enlarged view of raft

7 × 12 km showing 8° rotation anticlockwise, causing the clear lane downstream of island 'I' to be curved. Leads 'L' downstream of the crater and small island at lower right are almost straight, indicating unidirectional drift slightly north of westward. Note pressure ridges 'P' upstream of islands. Arrows show relative motion vectors of individual plates. Scale bar is 10 km.

discharge rates³ of the order of $10^6 \text{ m}^3 \text{ s}^{-2}$, causing a flood rich in suspended sediment of all sizes.

Water evaporation would be rapid under present Mars conditions¹⁹, but early work⁹ indicated that freezing rates at the surface of martian lakes would be of the order of 10^{-5} to $10^{-4} \text{ cm s}^{-1}$, and that surface ice will grow to a thickness of 5–10 m after 1 yr. Freezing will continue until a depth of 50 m is frozen solid in 5–10 yr. Recent work emphasizes that the water should have high concentrations of dissolved salts^{20,21}, and if it originates from magmatic intrusion, could be several degrees to tens of degrees above freezing on emergence³. These factors could produce longer timescales for complete freezing to occur¹⁰, and rapid surface heat loss could

cause intense convection that would prevent surface ice formation in the early stages, but allow slush to form.

Ice is unstable at the surface of Mars at present owing to sublimation in the 6-mbar atmosphere, but it is thought that huge volumes of volcanic ash also erupted from Cerberus Fossae⁶, which—if contemporaneous with water emission—would have formed a substantial protective layer²² on the ice. Depending on the porosity and thermal properties of this layer, the subsequent lowering of the floe surfaces could be very slow²³. Sublimating water vapour migrating through the pores will help over time to sinter and chemically bind the particles to form a stronger sublimation lag. To account for the 1-Myr age difference between the plates (pack-ice floes) and the younger lanes in between, we suggest the following sequence of events: first, pack-ice formation with a volcanic ash covering, second, remobilization, break-up and drift of pack-ice, with cessation of volcanic activity, third, freezing of the entire body of water, and finally, the sublimation of the unprotected ice between the ash-covered ice-floes, gradually exposing the suspended sediment at the surface to form a protective layer²² with a younger age than the floes. Alternatively, the latest Athabasca volcanic activity may be as young as 3 Myr⁵, which may have scattered ash that prevented further sublimation of inter-floe areas at this later time. This interpretation is supported by the MOLA profiles, which show them to be up to 3 m lower than the floes.

We do not know whether the frozen body of water is still there, or whether the visible floes are preserved in a sublimation residue draped over the substrate. Two observations suggest that it is still there:

(1) MOLA profiles show that three submerged or partially submerged craters 1.8 to 4.8 km in diameter have depths 2% to 3% of their diameter (Fig. 4), whereas the mean depth of a martian crater of this size range is about 20% of its diameter¹⁶, suggesting that most of the ice is still within the crater, though up to 15% by volume of the crater filling may be suspended sediment. Other submerged craters appear to have similar depths. This point depends on the craters being fresh rather than degraded before

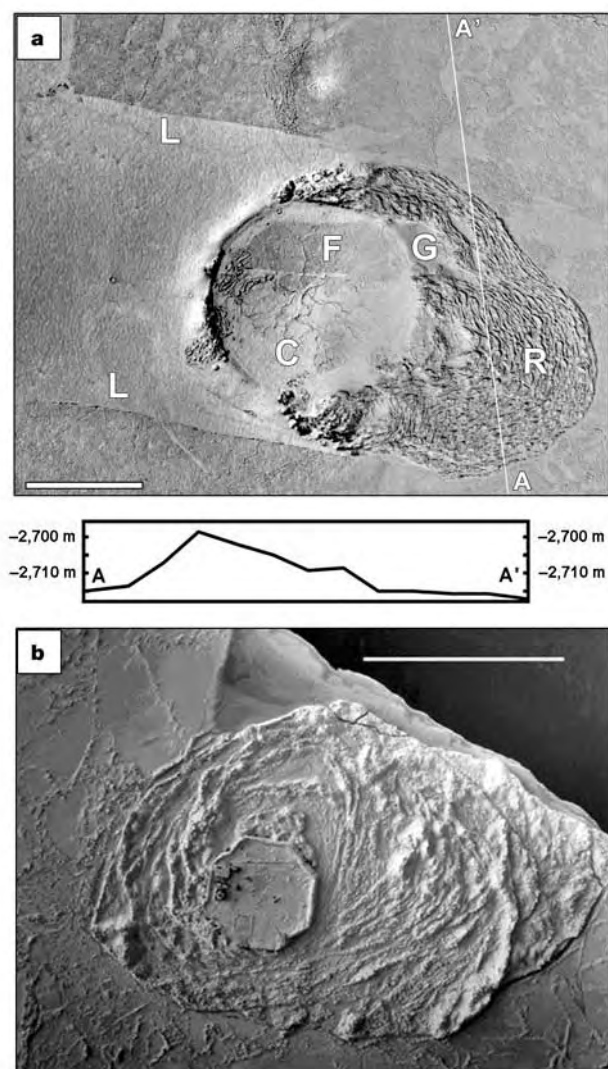


Figure 2 Pressure ridges on Mars, and those caused on Earth by pack-ice drift against obstacles. **a**, High-resolution MOC images E2100112 and R0900475 of an impact crater seen in Fig. 1a. The sides 'L' of the downstream lane are smooth at the 10-m scale, and parallel. The line on the right indicates the MOLA track 19960 crossing the pressure ridges, which are up to 16 m above the surrounding level. The MOLA topographic profile (values in metres) is shown below the image ($\times 60$ vertical exaggeration). Scale bar is 1 km. **b**, Air photo of Tarsiut Island, an artificial island in the Beaufort Sea, surrounded by first-year sea ice 1 to 1.5 m thick in the winter of 1983. We note 5-m-high concentric pressure ridging with wavelengths of 2 to 30 m caused by pack-ice rubble becoming grounded on the sloping flanks of the island during winds from different directions. Scale bar is 100 m. (Gulf Canada Resources photo courtesy of T.J.O. Sanderson.) 'C', surface cracking; 'F', a plate that has drifted through a gap 'G' in the rim; 'R', pressure ridges.

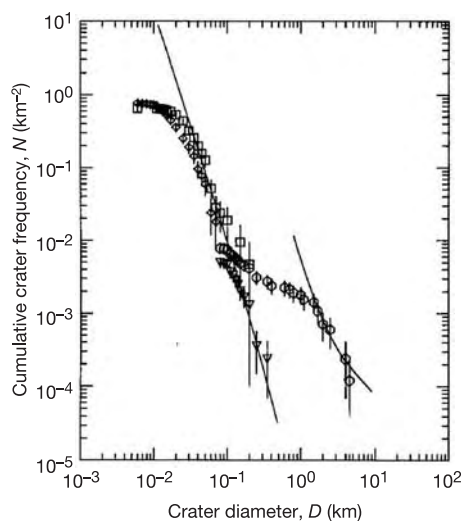


Figure 3 Age dating by crater counting^{12,13} on the pack-ice surface using HRSC (triangles) and MOC imagery over a total area of 380 km^2 (squares and diamonds), including those craters that protrude through the surface from the substratum (circles). Whereas the HRSC data indicate a single re-surfacing event about 5 ± 2 million years ago, counts on MOC images show consistently lower numbers for the brighter inter-plate terrain (diamonds) than the darker plate-like terrain (squares) interpreted as ice floes. The lower frequency occurs at all sizes systematically, indicating an age difference of about 1 Myr. The derived age of the substrate before flooding is 3.66 ± 0.05 billion years.

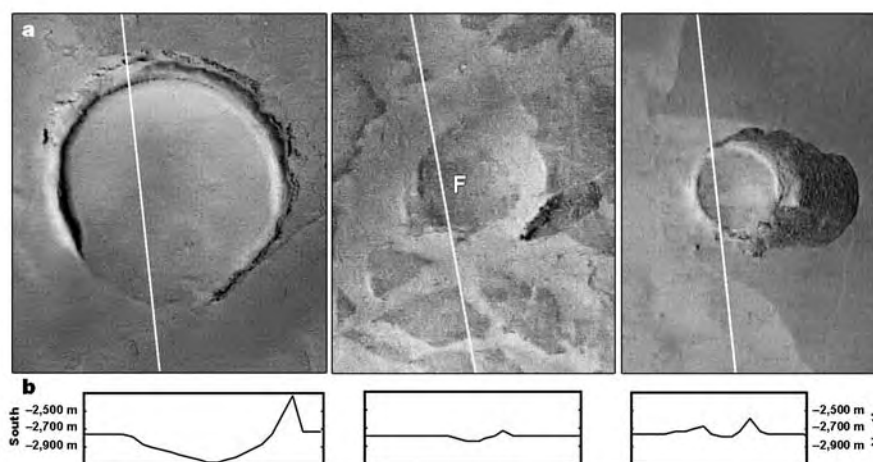


Figure 4 Evidence of ice surface lowering and draping of plate-like features over partly submerged impact craters. MOLA tracks (**a**) and MOLA topographic profiles with 10 × vertical exaggeration (**b**) across three flooded or partially flooded craters 4.8, 2.3 and 1.8 km in diameter from left to right. Values are in metres. Note that the floor sags at least

85, 18 and 25 m respectively, suggesting that 89% to 96% of the proposed ice fill is still within the crater. Note also the raft 'F' draped over the rim and floor of the centre crater. See text for further details.

flooding, but in this area most of the unflooded craters of similar size appear to be fresh with bowl-shaped interiors.

(2) MOLA profiles show a virtually horizontal surface, whereas the ice depth estimates above indicate that the substrate varies in altitude by 55 m. If the ice had been lost, sediment draped over this should have resulted in considerable surface height variation.

Recent work^{24,25} has shown that Mars' obliquity, with oscillations of 5° to 10° amplitude and periods of 10⁵ yr, was 37° ± 7° between 5 and 10 Myr ago. Global climate model simulations indicate that this would produce a substantially different climate from that of today, with higher dust transport and an atmosphere of higher temperature and pressure²⁶. The extremely young age of 5 Myr for the flood suggests that catastrophic flood events from a proposed sub-cryospheric aquifer^{20,21} are continuing to happen, as they have done throughout the known history of Mars' surface. The continuous presence of warm water beneath the cryosphere over several billion years might provide more opportunities for life to develop than was once thought. Microorganisms found within deep-sea hydrothermal vent communities²⁷ are common ancestors to many forms of life on Earth, and the possibility of life developing at similar places elsewhere in the Solar System has been postulated²⁸. □

Received 3 September 2004; accepted 18 January 2005; doi:10.1038/nature03379.

- Berman, D. C. & Hartmann, W. K. Recent fluvial, volcanic and tectonic activity on the Cerberus Plains of Mars. *Icarus* **159**, 1–17 (2002).
- Burr, D. M., Grier, J. A., McEwen, A. S. & Keszthelyi, L. P. Repeated aqueous flooding from the Cerberus Fossae: Evidence for very recently extant, deep groundwater on Mars. *Icarus* **159**, 53–73 (2002).
- Head, J. W. III, Wilson, L. & Mitchell, K. L. Generation of recent massive water floods at Cerberus Fossae, Mars, by dike emplacement, cryospheric cracking, and confined aquifer groundwater release. *Geophys. Res. Lett.* **30**, 1577, doi:10.1029/2003GL017135 (2003).
- Plescia, J. B. Cerberus Fossae, Elysium, Mars: a source for lava and water. *Icarus* **164**, 79–95 (2003).
- Werner, S. C., van Gasselt, S. & Neukum, G. Continual geological activity in Athabasca Valles Mars. *J. Geophys. Res.* **108**, E12, 8081, doi:10.1029/2002JE002020 (2003).
- Keszthelyi, L., McEwen, A. S. & Thordarson, T. Terrestrial analogs and thermal models for Martian flood lavas. *J. Geophys. Res.* **105**, 15027–15049 (2000).
- Fuller, E. R. & Head, J. W. III Amazonis Planitia: the role of geologically recent volcanism and sedimentation in the formation of the smoothest plains on Mars. *J. Geophys. Res.* **107**, 5081, doi:10.1029/2002JE001842 (2002).
- Lanagan, P. D. & McEwen, A. S. Cerberus Plains volcanism: constraints on temporal emplacement of the youngest flood lavas on Mars. *6th Int. Conf. on Mars* abstr. 3215 (2003); (<http://www.lpi.usra.edu/meetings/sixthmars2003/pdf/3215.pdf>).
- Carr, M. H. Stability of Martian streams and lakes. *Icarus* **56**, 476–495 (1983).
- Kreslavsky, M. A. & Head, J. W. III Fate of outflow channel effluents in the northern lowlands of Mars: the Vastitas Borealis formation as a sublimation residue from frozen ponded bodies of water. *J. Geophys. Res.* **107**, 5121, doi:10.1029/2001JE001831 (2001).

- Carr, M. H. & Head, J. W. III Oceans on Mars: an assessment of the observational evidence and possible fate. *J. Geophys. Res.* **108**, 5042, doi:10.1029/2002JE001963 (2003).
- Hartmann, W. K. & Neukum, G. Cratering chronology and the evolution of Mars. *Space Sci. Rev.* **96**, 165–194 (2001).
- Neukum, G. *et al.* Recent and episodic volcanic and glacial activity on Mars revealed by the High Resolution Stereo Camera. *Nature* **432**, 971–979 (2004).
- Patrick, M. R., Dehn, J. & Dean, K. Numerical modelling of lava flow cooling applied to the 1997 Okmok eruption: approach and analysis. *J. Geophys. Res.* **109**, B03202, doi:10.1029/2003JB002537 (2004).
- Skinner, B. J. Thermal expansion. In *Handbook Of Physical Constants* (ed. Clark, S. P.) *Geol. Soc. Am. Mem.* **97**, 75–96 (1966).
- Garvin, J. B., Sakimoto, S. E. H. & Frawley, J. J. Craters on Mars: global geometric properties from gridded MOLA topography. *Sixth Int. Conf. on Mars* abstr. 3277 (2003); (<http://www.lpi.usra.edu/meetings/sixthmars2003/pdf/3277.pdf>).
- Rice, J. W., Parker, T. J., Russel, A. J. & Knudsen, O. Morphology of fresh outflow channel deposits on Mars. *Lunar Planet. Sci. XXXIII* abstr. 2026 [CD-Rom] (Lunar & Planetary Institute, Houston, Texas, 2002).
- Baker, V. R. *et al.* in *Mars* (eds Kieffer, H. H., Jakosky, B. M., Snyder, C. W. & Matthews, M. S.) 493–522 (Univ. Arizona Press, Tucson, 1992).
- Wallace, D. & Sagan, C. Evaporation of ice in planetary atmospheres: ice-covered rivers on Mars. *Icarus* **39**, 476–495 (1983).
- Clifford, S. M. A model for the hydrologic and climatic behavior of water on Mars. *J. Geophys. Res.* **98**, 10973–11016 (1993).
- Clifford, S. M. & Parker, T. J. The evolution of the Martian hydrosphere: implications for the fate of a primordial ocean and the current state of the northern plains. *Icarus* **154**, 40–79 (2001).
- Carr, M. H. D/H on Mars: effects of floods, volcanism, impacts and polar processes. *Icarus* **87**, 210–227 (1990).
- Skorov, V. V., Markiewicz, W. J., Basilevsky, A. T. & Keller, H. U. Stability of water ice under a porous nonvolatile layer: implications for the south polar layered deposits of Mars. *Planet. Space Sci.* **49** (1), 59–63 (2001).
- Laskar, J. & Robutel, L. The chaotic obliquity of the planets. *Nature* **361**, 608–612 (1993).
- Laskar, J., Gastineau, M., Joutel, F., Levrard, B. & Robutel, P. A new astronomical solution for the long term evolution of the insolation quantities of Mars. *Lunar Planet. Sci. XXXV* abstr. 1600 (2004).
- Haberle, R. M., Murphy, J. R. & Schaeffer, J. Orbital change experiments with a Mars general circulation model. *Icarus* **161**, 66–89 (2003).
- Humphris, S. E., Zierenberg, R. A., Mullineaux, L. S. & Thomson, R. E. (eds) *Seafloor Hydrothermal Systems* 1–466 (AGU Geophys. Monogr. 91, American Geophysical Union, Washington, DC, 1995).
- Greenberg, R., Geissler, P., Tufts, B. R. & Hoppa, G. V. Habitability of Europa's crust: The role of tidal-tectonic processes. *J. Geophys. Res.* **105**, 17551–17562 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M. Wählisch for assistance in the MOLA processing and S. Clifford and N. A. Cabrol for criticism that greatly improved the paper.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.B.M. (J.B.Murray@open.ac.uk).

The HRSC Co-Investigator Team: J. Albertz¹, A. T. Basilevsky², G. Bellucci³, J.-P. Bibring⁴, M. Buchroithner⁵, M. H. Carr⁶, E. Dorrer⁷, T. C. Duxbury⁸, H. Ebner⁹, B. H. Foing¹⁰, R. Greeley¹¹, E. Hauber¹², J. W. Head III¹³, C. Heipke¹⁴, H. Hoffmann¹⁵, A. Inada¹⁶, W.-H. Ip¹⁷, B. A. Ivanov¹⁸, R. Jaumann¹⁹, H. U. Keller²⁰, R. Kirk²¹, K. Kraus²², P. Kronberg²³, R. Kuzmin²⁴, Y. Langevin²⁵, L. Lumme²⁶, W. Markiewicz²⁷, P. Masson²⁸, H. Mayer²⁹, T. B. McCord³⁰, J.-P. Muller³¹, J. B. Murray³², F. M. Neubauer³³, G. Neukum (PI)²⁸, J. Oberst¹², G. G. Ori²⁹, M. Pätzold²⁷, P. Pinet³⁰, R. Pischel¹², F. Poulet⁴, J. Raitala³¹, G. Schwarz³², T. Spohn¹² & S. W. Squyres³³

Affiliations of authors: 1, TU Berlin, D-10623 Germany; 2, Vernadsky Institute-RAS, Moscow 117975, Russia; 3, IFSI/CNR, Rome, Italy; 4, IAS, Orsay Campus 91405, France; 5, TU Dresden D-01062, Germany; 6, USGS, MS 975 Menlo Park, California 94025, USA; 7, Universität der Bundeswehr München D-85577, Germany; 8, JPL, Pasadena, California 91109, USA; 9, TU München, D-80333 Germany; 10, ESTEC/SCI-SR, Postbus 299, NL-2200 AG Noordwijk, The Netherlands; 11, Arizona State University, Box 871404, Tempe, Arizona 85287-1404, USA; 12, DLR, D-12489 Berlin, Germany; 13, Brown University, Box 1846, Providence, Rhode Island 02912, USA; 14, Universität Hannover, D-30167 Germany; 15, Kobe University, Kobe, Japan; 16, National Central University (NCU) Taiwan; 17, IDG-RAS, Moscow 117979, Russia; 18, MPAE, Postfach 20, D-37191 Katlenburg-Lindau, Germany; 19, USGS, Flagstaff, Arizona 86001, USA; 20, TU Wien, A-1040 Wien, Austria; 21, TU Clausthal, D-38678 Germany; 22, University of Helsinki, PO Box 14, SF-00014 Helsinki, Finland; 23, OrsayTerre, F-91405 Orsay Campus, France; 24, PSI-Nw, Winthrop, Washington 98862, USA; 25, University College London, WC1E 6BT, UK; 26, The Open University, Milton Keynes MK7 6AA, UK; 27, Universität Köln, D-50923 Germany; 28, FU Berlin, D-12249 Germany; 29, IRSPS, I-65127 Pescara, Italy; 30, Observatoire de Midi-Pyrénées, F-31400 Toulouse, France; 31, University of Oulu, Fin-90401 Finland; 32, DLR, D-82234 Wessling, Germany; 33, Cornell University, Ithaca, New York 14853-1301, USA

Discovery of a flank caldera and very young glacial activity at Hecates Tholus, Mars

Ernst Hauber¹, Stephan van Gassel², Boris Ivanov³, Stephanie Werner², James W. Head⁴, Gerhard Neukum², Ralf Jaumann¹, Ronald Greeley⁵, Karl L. Mitchell⁶, Peter Muller⁷ & The HRSC Co-Investigator Team*

¹Institute of Planetary Research, German Aerospace Center (DLR), 12489 Berlin, Germany

²Institute of Geosciences, FU Berlin, 12249 Berlin, Germany

³Institute for Dynamics of Geospheres, Russian Academy of Sciences, Moscow, 117334, Russia

⁴Department of Geological Sciences, Brown University, Providence, Rhode Island 02912, USA

⁵Department of Geological Sciences, Arizona State University, Tempe, Arizona 85287-1404, USA

⁶Environmental Science Department, Lancaster University, Lancaster LA1 4YQ, UK

⁷Department of Geomatic Engineering, University College London, Gower Street, London WC1E 6BT, UK

*A list of all members of The HRSC Co-Investigator Team and their affiliations appears at the end of the paper

The majority of volcanic products on Mars are thought to be mafic and effusive^{1,2}. Explosive eruptions of basic to ultrabasic chemistry are expected to be common^{3,4}, but evidence for them is rare and mostly confined to very old surface features⁵. Here we present new image and topographic data from the High Resolution Stereo Camera that reveal previously unknown traces of an explosive eruption at 30° N and 149° E on the northwestern flank of the shield volcano Hecates Tholus. The eruption created

a large, 10-km-diameter caldera ~350 million years ago. We interpret these observations to mean that large-scale explosive volcanism on Mars was not confined to the planet's early evolution. We also show that glacial deposits partly fill the caldera and an adjacent depression. Their age, derived from crater counts, is about 5 to 24 million years. Climate models predict that near-surface ice is not stable at mid-latitudes today⁶, assuming a thermo-dynamic steady state. Therefore, the discovery of very young glacial features at Hecates Tholus suggests recent climate changes. We show that the absolute ages of these very recent glacial deposits correspond very well to a period of increased obliquity of the planet's rotational axis⁷.

The ESA Mars Express mission, an orbiter carrying seven experiments, was inserted into Mars orbit on 25 December 2003. On 19 January 2004, the multiple line scanner instrument, the High Resolution Stereo Camera (HRSC)⁸, imaged the volcano Hecates Tholus in the Elysium region. Our study focuses on two overlapping depressions at the northwestern base of Hecates Tholus (Fig. 1) that were mentioned before⁹, but without an explanation for their origin. The HRSC image resolution of that area (~26 m per pixel) is better than that of previous images from the Viking Orbiter camera (~40 m per pixel) and from the THEMIS thermal infrared imager (~100 m per pixel). Several very high-resolution images from the Mars Orbiter Camera (MOC) cover small parts of the depressions with 3 to 4 m per pixel. We use digital photogrammetric techniques¹⁰ to derive stereo information with a mean relative point accuracy of ~30 m from the HRSC's multiple line sensors, which observe the surface under different viewing angles.

The smaller of the two depressions (here referred to as 'depression A') has an area of ~12 km × 10 km (Fig. 2a) and a depth between 1,000 and 1,500 m. The northwestern part of its rim is missing where it overlaps with the larger depression (here named 'depression B'). The remaining rim has an elevation between 800 and 1,800 m. The floor is terraced, with an elevation difference of 200–300 m between the two levels. Owing to the incomplete rim, it is difficult to determine its volume. Our best estimate, based on a reconstructed rim, is ~80 km³. On the flanks of the volcano, an unusual hilly and knobby deposit can be distinguished adjacent to depression A. Its surface is rougher than the rest of the flank's surface, and it extends outward from the rim to a maximum distance of about 15 km.

We favour a volcanic over an impact origin of depression A for four reasons. First, the morphology of the depression, including the two different levels of its floor, is remarkably similar to part of the caldera complex at the shield volcano Ascraeus Mons in the Tharsis region (Fig. 2b), and also to the summit caldera of Hecates Tholus itself (Fig. 2c); impact craters on Hecates Tholus have a distinctly different appearance (Fig. 2d). Second, the stereo information indicates that the walls slope at an average angle of about ~30°, which is steeper than the walls of most martian impact craters¹¹. Third, there is no elevated crater rim, which would be expected if depression A were an impact crater. Fourth, the remaining parts of the rim are distinctly not circular, owing to a promontory at the topographically highest part of the rim.

Hence, the cumulative evidence of these independent observations suggests that depression A is volcanic rather than impact-related. There is no evidence for effusive eruptions, for example, lava flows, near depression A. Instead, we interpret the rough material near depression A as the proximal part of pyroclastic materials from an explosive eruption. Relative to the other parts of the flanks, an area between depression A and the summit caldera displays a lack of impact craters and a generally smooth surface texture at the scale of the Viking and HRSC image resolution. It has been interpreted to be a mantling deposit from an explosive eruption at the summit⁹. However, it may as easily have been produced by an explosion at depression A. Indeed, the isolines of the crater density on the western flank of Hecates Tholus (figure 7 in ref. 9) are roughly

concentric around depression A and would be in better agreement with an explosion there than with one at the summit. We interpret the smooth material as the distal part of the erupted pyroclastic material. The presence of many fluvial channels^{9,12} may indicate

phreatomagmatic (containing magmatic gases and steam) interactions, which could have enhanced the explosivity of the eruption. Crater counts on HRSC and MOC images on both the proximal and distal pyroclastic material, using a new model of cratering

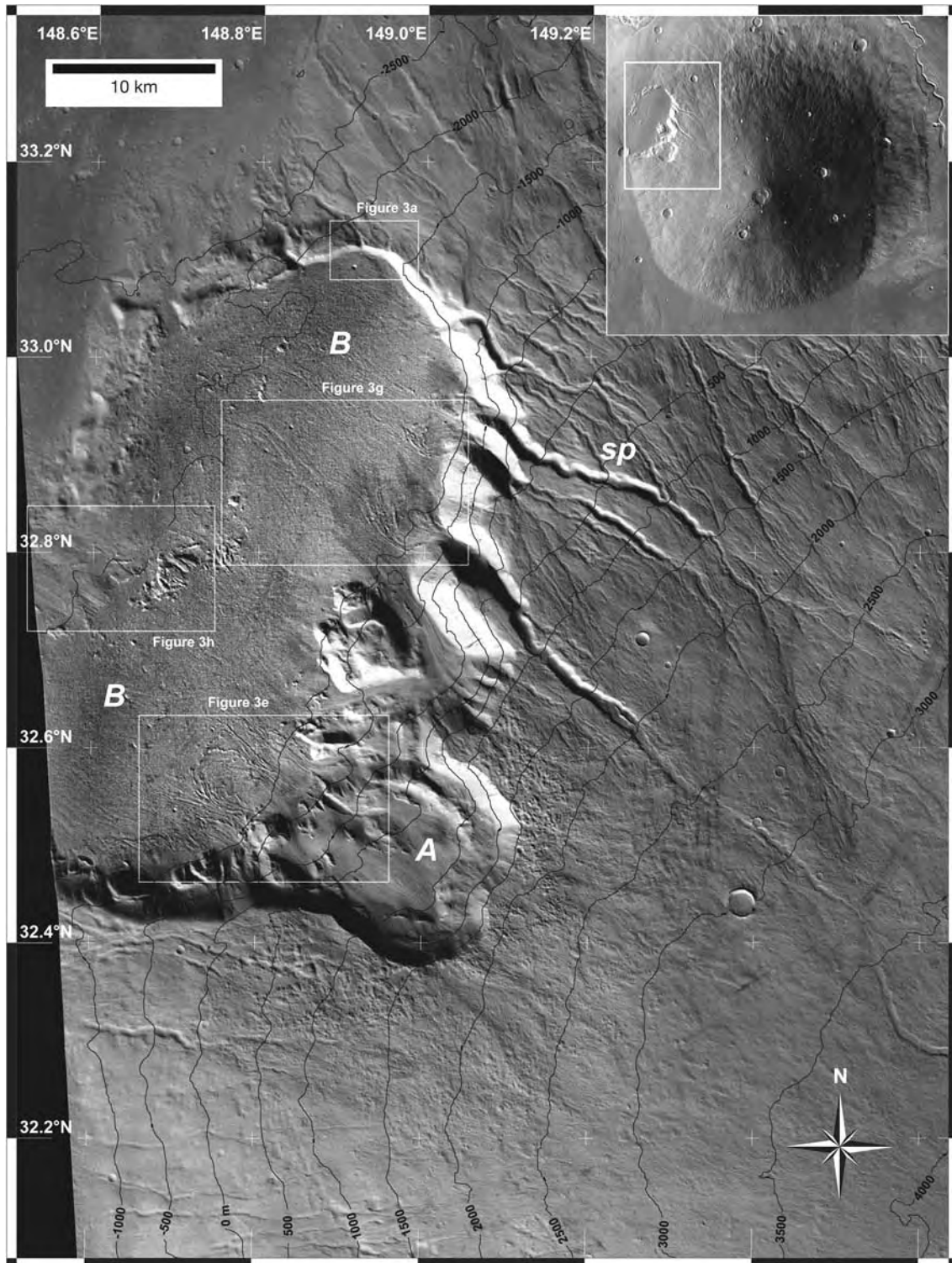


Figure 1 Topographic image map of the study area at the base of the northwestern flank of Hecates Tholus (part of HRSC image h0032_0000.nd). Topographic information (contour line distance 500 m; reference plane is the Mars IAU 2000 ellipsoid) was derived from HRSC stereo imagery. The smaller depression (A) is surrounded by a unique knobby and hilly material. Note the stream piracy pattern (sp) immediately east of a larger

depression (B), indicating that an older trend of flow directions was changed as a result of the formation of depression B. The white boxes show the locations of Fig. 3a, e, g and h. The inset on the upper right is a mosaic of Themis-infrared daytime images and shows the location of the base map.

chronology^{13,14}, give an absolute age for the eruption of ~350 Myr. Hence, large-scale explosive volcanism occurred in the last 10% of the planet's history. This is very young when compared to the several-billion-year-old shields in the highlands, which are among the best-documented examples yet of explosive volcanism on Mars⁵.

The shape and distribution of channels near the caldera exhibit several characteristics that shed light on the chronology of volcanic and fluvial processes. East of depression B, we observe two peaks in the azimuthal distribution of the channels. A first set of channels (set A) has an orientation of about N 30° W and is cut off by depression B at several locations. These channels bifurcate and meander where minor surface undulations cause a decrease in flow energy. Several channels that are cut off by depression B can be traced towards the north of the depression, where they seem to continue on its rim. In at least one example (Fig. 3a), the base of a channel starts at a topographically higher level than the floor of depression B, indicating that the channel is older.

A second, younger set of channels (set B) with an azimuthal trend of N 50° W to N 70° W deviates from and partly crosses the older set A, creating a stream piracy pattern. Its channels are deeper and broader than those of set A, and groundwater sapping from a water-rich subsurface might have contributed to their morphology. Set B starts several kilometres away from the eastern rim of depression B, where the topographic gradient becomes higher and is directed towards its rim. The channels follow this topographic trend and deposited large amounts of debris onto the floor. These observations suggest that fluvial activity on Hecates Tholus was continual, not episodic, during the events which formed the depressions, and that the interaction of magma and water or ice may have contributed to the explosive nature of the eruption.

The lower level of depression A and several smaller valleys near the walls of depression B are covered by a smooth deposit, which is lineated in a downslope direction (Fig. 3b). It resembles the lineated valley fill in the fretted terrain near the dichotomy boundary, which has been interpreted as rock glaciers^{15,16}. Where the lineated material flows over a topographic step, its surface is distinctively rougher than on flat ground (Fig. 3c). This pattern resembles the change in surface texture encountered at terrestrial icefalls (Fig. 3d). Beyond the topographic step, it extends outward onto the floor of depression B for a distance of ~6 km. It is bounded by curvilinear ridges that resemble terrestrial end moraines (Fig. 3e). Lobate flow features also extend away from the base

of the wall for ~2.5 km (Fig. 3f). Where several valleys debouch into depression B, fan deposits extend for up to 6 km on the surrounding plains (Fig. 3g). We interpret them as debris, transported down the strongly incised channels of set B. Alternatively, these deposits could also be moraines. Long and slightly sinuous features extend downslope (~1° slope) from the end of the fan deposits across the entire floor of depression B towards the northwest. We interpret them as distal meltwater channels extending across a proglacial braided outwash plain, analogous to an Icelandic sandur.

A topographic ridge separates depression B from the topographically lower lava flows from Elysium Mons, which are located further towards the northwest. Where this ridge is breached, some rounded, low and shallow hills are superposed by straight, long and narrow ridges and trenches (Fig. 3h). They resemble terrestrial subglacial erosion features (for example, drumlins or whalebacks), and indicate that the glaciation possibly extended beyond depression B towards the northwest. As elsewhere on Mars¹⁷, the strongest arguments for a glacial origin are the assemblage of various surface features that are strikingly similar to terrestrial glacial landforms (medial moraines, end moraines, meltwater channels, drumlins) and their consistently glacial proximal-to-distal relationship (Figs 1 and 3). The volatile most likely to have formed the glaciers is water-ice¹⁷, as the only alternative, CO₂, is particularly unstable at low latitudes under any conceivable atmospheric conditions. The water source could have been precipitation or groundwater that freezes when coming into contact with ice. Precipitation of water on the martian surface is known to take place even under the current thin atmosphere¹⁸. There are no obvious surficial pathways of water into depression A. Hence, groundwater emerging at the base of scarps bounding the depressions might have fed the glaciers in depression A. However, the local microenvironment at the floors of the depressions, which are partly protected from insolation by steep walls, might act as a cold trap to enhance frost deposition as a water source.

We performed crater counts on HRSC and MOC images of the glacial features shown in Fig. 3, and obtained cratering model¹⁴ ages between ~25 and 5 Myr (Fig. 4). The detection of very young glacial features at a latitude of 30° N in the Elysium region has profound implications for the recent martian climate history. Geologic observations suggest that Mars has experienced recent ice ages^{19,20}, and it is inferred from gamma-ray and neutron spectroscopy that water might indeed be present today near the surface²¹. However,

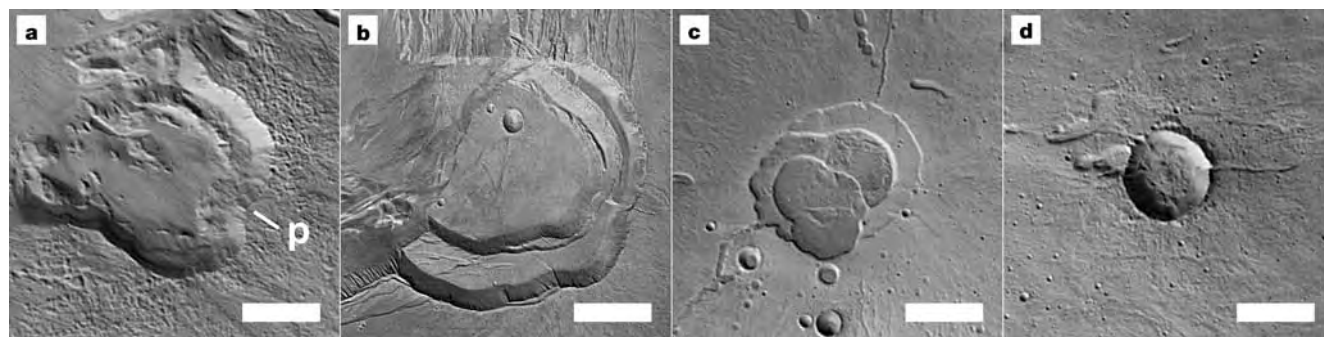


Figure 2 Morphology of calderas. **a**, Feature interpreted to be a caldera on the northwestern flank of Hecates Tholus (depression A of Fig. 1). Owing to a promontory (p), the outline is distinctly non-circular, making an impact origin improbable. Note the two levels of floor elevation, with lineations in the SE–NW direction filling the lower level. **b**, Part of the summit caldera complex of Ascraeus Mons (part of HRSC image h0068_0000.nd, taken on 31 January 2004; centre at 11.13° N, 255.95° E). Note the two levels of floor elevation, similar to what is observed in depression A. **c**, Summit caldera

of Hecates Tholus (32.05° N, 150.1° E). Note the morphologic similarity of this caldera and the calderas shown in **a** and **b**. **d**, The largest impact crater on Hecates Tholus (32.28° N, 150.8° E) displays a sharp, elevated crater rim and a continuous floor that is not divided into two elevation levels. Note the overall dissimilarity with **a**, **b** and **c**. In all images, North is up and the white scale bar is 5 km. Figures 2a, c and d are details of HRSC h0032_0000.nd.

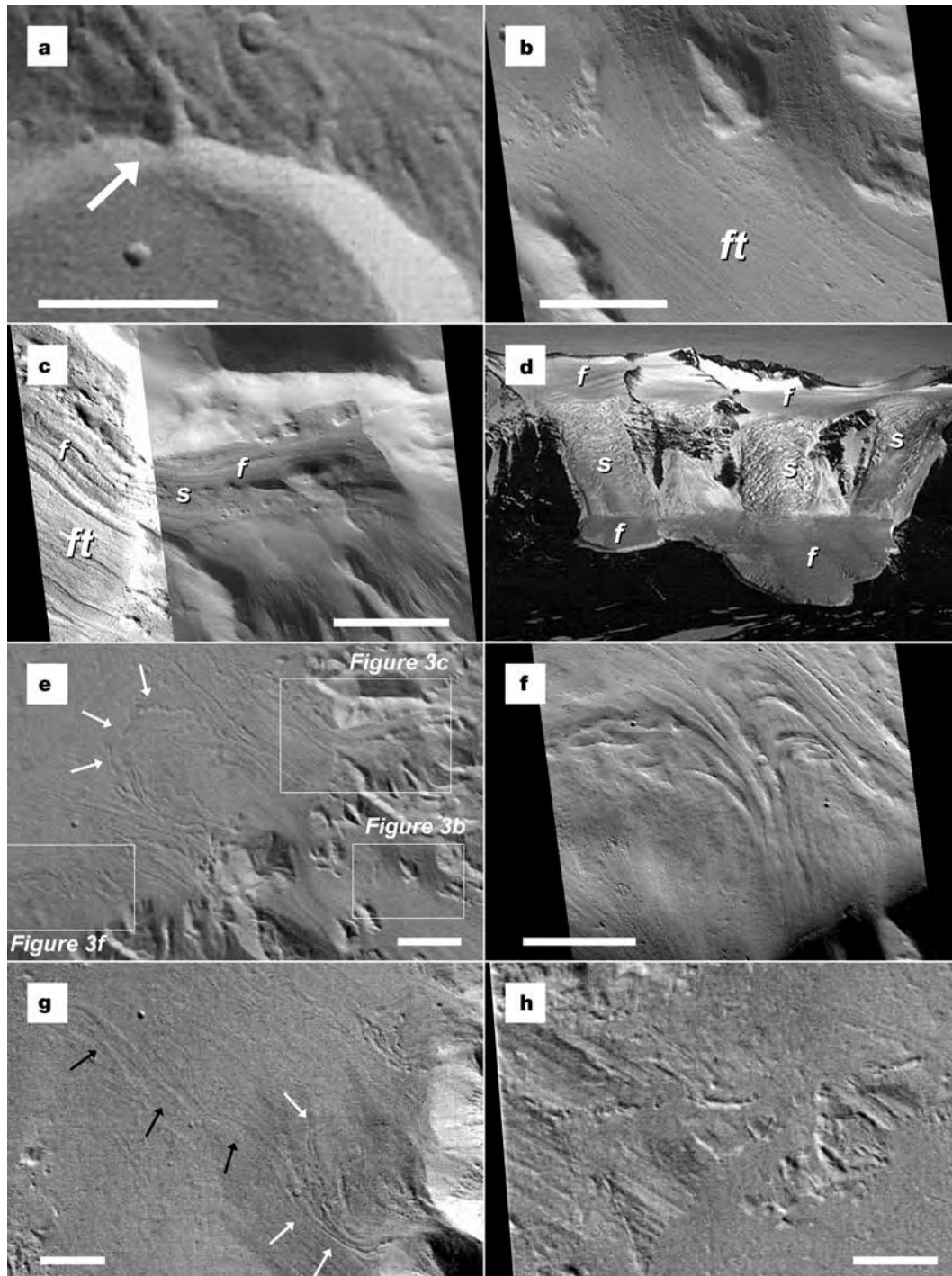


Figure 3 Surface landforms indicative of fluvial and glacial processes at the northwestern flank of Hecates Tholus. **a**, A valley cut upstream by depression B resembles a hanging valley, although true hanging valleys are cut downstream by other valleys. The valley predates the depression (see Fig. 1 for location). **b**, Lineated material (ft), resembling terrestrial medial moraines. A similar surface texture is observed at the fretted terrain^{15,16} (MOC image R0802750; see Fig. 3e for location). **c**, Lineated material flowing over the topographic step between depressions A (higher, left) and B (lower, right). Where the topographic gradient is steep (s), the surface texture is rougher than on flat terrain (f), resembling the change in texture observed in terrestrial icefalls (compare with Fig. 3d; MOC images M03-01763 and R08-02750; see Fig. 3e for location). **d**, Icefalls in Taylor Valley, Antarctica (77° 45' S, 162° 30' E). Note the change in surface texture between flat

(f) and steep (s) terrain, similar to Fig. 3c. Photo courtesy of T. Lowell. **e**, Curvilinear, moraine-like features (white arrows) downslope of the topographic scarp between depressions A and B (see Fig. 1 for location; white boxes show location of Figs 3b, c and f). **f**, Lobate flow features near the base of the wall of depression B, resembling glacial flow features on Earth (MOC image R09-04137; see Fig. 3e for location). **g**, Fan deposits (white arrows) in depression B at the terminations of deeply incised valleys. Faint and slightly sinuous features (black arrows) indicate water runoff from the deposits (see Fig. 1 for location). **h**, Low, rounded hills with superposed long and thin ridges, similar to terrestrial subglacial erosion features (see Fig. 1 for location). Scale bars in Fig. 3a, e, g and h (all part of HRSC image h0032_0000.nd) are 2 km, and in Fig. 3b, c and f are 1 km.

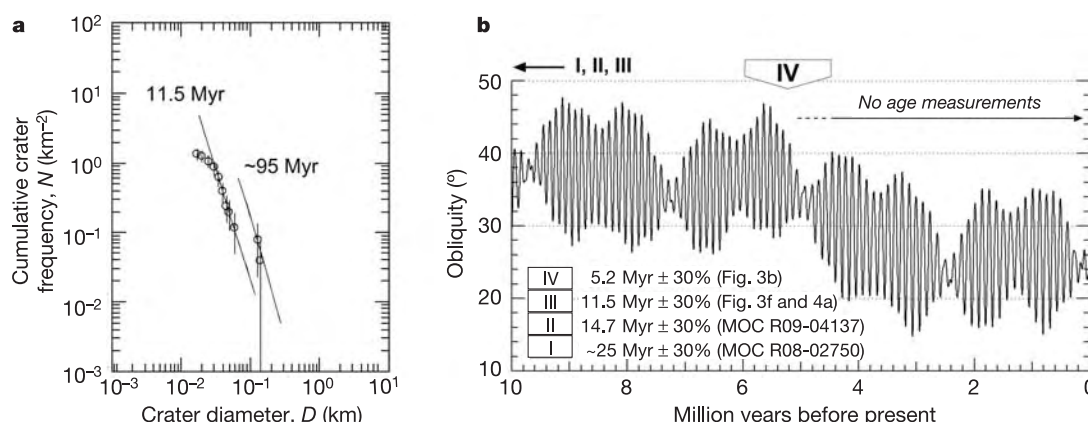


Figure 4 Chronology of glacial surface features and correlation to obliquity changes. **a**, Crater size-frequency distribution for the floor of depression B as observed in MOC R09-04137. Crater model curves according to a new cratering chronology model¹³ are fitted to the measurements. The crater model age is 11.5 Myr. An older crater population (right curve) corresponds to an age of ~95 Myr, which might be associated with an older episode of glaciation. Additional ages of 5.2, 14.7 and ~25 Myr were obtained from other MOC images (Fig. 4b). The different ages are in agreement with the locations of the geologic units, with the oldest glacial surfaces at the largest distance from the southeastern part of the depressions, and the youngest surface in the interior of depression A. The error inherent to this technique is $\pm 30\%$ (vertical black lines are error

bars). **b**, Changes in the obliquity of Mars' rotational axis over the last 10 Myr (ref. 7). The past ~10 Myr can be reliably modelled despite the chaotic behaviour of the orbits in the Solar System^{7,24}, so the large increase in obliquity at ~5 Myr is robust. The youngest crater model age derived from our crater counts is plotted as a broad arrow on top of the figure (the other ages fall outside the range shown here). The width of the arrow corresponds to the 30% uncertainty. Note that all our ages are older than the abrupt increase of obliquity at 5 Myr. Our age measurements suggest that the mean obliquity might have been higher than today even before 10 Myr ago, allowing ice to be present globally according to climate models⁶.

although there is a hydrogen-rich zone at Elysium²², the spatial resolution of these measurements is far too low to detect any local enrichment on the scale of the landforms seen at Hecates Tholus. At present, the pressure and temperature conditions of the martian atmosphere prevent near-surface ice from being stable at equatorial latitudes⁶. However, the obliquity of the planet's rotational axis varied significantly over the past 20 Myr (ref. 7). In periods of higher obliquity the climate was different from today's²³.

According to recent climate models, north polar water-ice could be mobilized under such conditions²⁴ and be deposited at mid- and low latitudes^{25,26} where it would have been stable⁶. This study is the first to combine calculations of orbital variations and climate models with the absolute dating of glacial surface features. The ages of glacial deposits on Hecates Tholus range from 25 to 5 Myr before present, with a $\pm 30\%$ error (Fig. 4a). This corresponds very well to a period of increased obliquity, which ended about 5 Myr ago⁷ (Fig. 4b). The averaged long-term obliquity between 5 and 20 Myr ago is $>35^\circ$, a value that is predicted by models to allow ice to be stable globally⁶. Hence, our observations show that the independent results on orbital variations⁷ and climate modelling^{24–26} are in chronological agreement with geologic surface features.

There are several reasons why ice may still be present at Hecates Tholus. The sublimation of ice results in the accumulation of sediment particles at the surface, and the formation of a lag deposit that is very effective in protecting ice from further sublimation²⁷. In addition, Elysium is a long-term sink of atmospheric dust²⁶, the deposition of which might have further decreased the sublimation rate. There is no evidence for significant degradation or for collapse features like kettle holes in images of the interior of depression A (Fig. 3b). We conclude that there may well have been some unknown amount of sublimation, but that ice is still buried and maintains the 'intact' appearance of the surface. On Earth, Miocene-aged ice (~8 Myr) is still present in the Antarctic dry valleys under a layer of sublimation till²⁸. Therefore, the ice at Hecates

Tholus may well have been preserved in very shallow depths for geologically long timescales^{29,30}, and could even be present today and accessible for automated or human exploration. □

Received 3 September 2004; accepted 2 February 2005; doi:10.1038/nature03423.

- Greeley, R. & Spudis, P. D. Volcanism on Mars. *Rev. Geophys. Space Phys.* **19**, 13–41 (1981).
- Mouginis-Mark, P. J., Wilson, L. & Zuber, M. T. in *Mars* (eds Kieffer, H. H., Jakosky, B. M., Snyder, C. W. & Matthews, M. S.) 424–452 (Univ. Arizona Press, Tucson/London, 1992).
- Francis, P. W. & Wood, C. A. Absence of silicic volcanism on Mars: Implications for crustal composition and volatile abundance. *J. Geophys. Res.* **87**, 9881–9889 (1982).
- Wilson, L. & Head, J. W. Mars: review and analysis of volcanic eruption theory and relationships to observed landforms. *Rev. Geophys.* **32**, 221–263 (1994).
- Greeley, R. & Crown, D. A. Volcanic geology of Tyrrhena Patera, Mars. *J. Geophys. Res.* **95**, 7133–7149 (1993).
- Mellon, M. T. & Jakosky, B. M. The distribution and behaviour of martian ground ice during past and present epochs. *J. Geophys. Res.* **100**, 11781–11799 (1995).
- Laskar, J. *et al.* Long term evolution and chaotic diffusion of the insolation quantities of Mars. *Icarus* **170**, 343–364 (2004).
- Neukum, G. *et al.* HRSC: The High Resolution Stereo Camera of Mars Express 17–35 (Report ESA-SP-1240, European Space Agency Publications Division, Noordwijk, The Netherlands, 2004).
- Mouginis-Mark, P. J., Wilson, L. & Head, J. W. Explosive Volcanism on Hecates Tholus, Mars: Investigation of Eruption Conditions. *J. Geophys. Res.* **87**, 9890–9904 (1982).
- Wewel, E., Scholten, F. & Gwinner, K. High Resolution Stereo Camera (HRSC)—Multispectral 3D-data acquisition and photogrammetric data processing. *Can. J. Remote Sens.* **26**, 466–474 (2000).
- Werner, S. C., Ivanov, B. A. & Neukum, G. Impact cratering on Mars: Search for target influence on morphology. *Lunar Planet. Sci.* **XXXV**, abstr. 1953 [CD-ROM] (Lunar and Planetary Institute, Houston, 2004).
- Gulick, V. C. & Baker, V. R. Origin and evolution of valleys on martian volcanoes. *J. Geophys. Res.* **95**, 14325–14344 (1990).
- Hartmann, W. K. & Neukum, G. Cratering chronology and the evolution of Mars. *Space Sci. Rev.* **96**, 165–194 (2001).
- Neukum, G. *et al.* Mars: Recent and episodic volcanic and glacial activity on Mars revealed by the High Resolution Stereo Camera. *Nature* **432**, 971–979 (2004).
- Squires, S. W. Martian fretted terrain: flow of erosional debris. *Icarus* **34**, 600–613 (1978).
- Mangold, N. Geomorphic analysis of lobate debris aprons on Mars at Mars Orbiter Camera scale: Evidence for ice sublimation initiated by fractures. *J. Geophys. Res.* **108**, 8021, doi:10.1029/2002JE001885 (2003).
- Baker, V. R. Water and the martian landscape. *Nature* **412**, 228–236 (2001).
- Wall, S. W. Analysis of condensates formed at the Viking 2 Lander site: The first winter. *Icarus* **47**, 173–183 (1981).
- Head, J. W. & Marchant, D. R. Cold-based mountain glaciers on Mars: Western Arsia Mons. *Geology* **31**, 641–644 (2003).
- Head, J. W., Mustard, J. F., Kreslavsky, A., Milliken, R. E. & Marchant, D. R. Recent ice ages on Mars. *Nature* **426**, 797–802 (2003).

21. Mitrofanov, I. *et al.* Maps of subsurface hydrogen from the high-energy neutron detector, Mars Odyssey. *Science* **297**, 78–81 (2002).
22. Feldman, W. C. *et al.* Global distribution of near-surface hydrogen on Mars. *J. Geophys. Res.* **109**, E09006, doi:10.1029/2003JE002160 (2004).
23. Jakosky, B. M., Henderson, B. G. & Mellon, M. T. Chaotic obliquity and the nature of the martian climate. *J. Geophys. Res.* **100**, 1579–1584 (1995).
24. Levrard, B., Forget, F., Montmessin, F. & Laskar, J. Recent ice-rich deposits formed at high latitudes on Mars by sublimation of unstable equatorial ice during low obliquity. *Nature* **431**, 1072–1075 (2004).
25. Mischina, M. A., Richardson, M. I., Wilson, R. J. & McCleese, D. J. On the orbital forcing of martian water and CO₂ cycles: A general circulation model study with simplified volatile schemes. *J. Geophys. Res.* **108**, 5062, doi:10.1029/2003JE002051 (2003).
26. Haberle, R. M., Murphy, J. R. & Schaeffer, J. Orbital change experiments with a Mars general circulation model. *Icarus* **161**, 66–89 (2003).
27. Skorov, Y. V. *et al.* Stability of water ice under a porous nonvolatile layer: implications to the south polar layered deposits of Mars. *Planet. Space Sci.* **49**, 59–63 (2001).
28. Marchant, D. R. *et al.* Formation of patterned ground and sublimation till over Miocene glacier ice in Beacon Valley, southern Victoria Land, Antarctica. *Geol. Soc. Am. Bull.* **114**, 718–730 (2002).
29. Mellon, M. T., Jakosky, B. M. & Postawko, S. E. The persistence of equatorial ground ice on Mars. *J. Geophys. Res.* **102**, 19357–19369 (1997).
30. Helbert, J. & Benkhoff, J. A new model on the thermal behavior of the near surface layer on Mars and its implications for ground ice deposits in Gusev Crater. *6th Int. Conf. Mars abstr.* 3019 [CD-ROM] (Lunar and Planetary Institute, Houston, 2003).

Acknowledgements We thank the entire HRSC experiment and instrument teams at the German Aerospace Center (DLR) and at the Freie Universität Berlin, as well as the Mars Express teams at ESTEC and ESOC. This study would not have been possible without their work. In particular, we appreciate the support of H. Hoffmann, T. Roatsch, K.-D. Matz, V. Mertens, J. Flohrer, R. Pischel, F. Scholten and K. Gwinner. HRSC was developed at DLR and industrial partners. G.N. is the Principal Investigator of this experiment. We are grateful to U. Wolf for her support in crater counting. M. Aittola, J. Korteniemi, P. Kostama and D. Williams supported the HRSC image planning. T. Lowell provided an image of his collection of glacier photographs. Comments by J. Helbert helped to improve the manuscript.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.H. (Ernst.Hauber@dlr.de).

The HRSC Co-Investigator Team: J. Alibert¹, A. T. Basilevsky², G. Bellucci³, J.-P. Bibring⁴, M. Buchroithner⁵, M. H. Carr⁶, E. Dorrer⁷, T. C. Duxbury⁸, H. Ebner⁹, B. H. Foing¹⁰, R. Greeley¹¹, E. Hauber¹², J. W. Head III¹³, C. Heipke¹⁴, H. Hoffmann¹², A. Inada^{15,18}, W.-H. Ip¹⁶, B. A. Ivanov¹⁷, R. Jaumann¹², H. U. Keller¹⁸, R. Kirk¹⁹, K. Kraus²⁰, P. Kronberg²¹, R. Kuzmin², Y. Langevin⁴, K. Lumme²², W. Markiewicz¹⁸, P. Masson²³, H. Mayer⁷, T. B. McCord²⁴, J.-P. Muller²⁵, J. B. Murray²⁶, F. M. Neubauer²⁷, G. Neukum²⁸, J. Oberst¹², G. G. Ori²⁹, M. Pätzold²⁷, P. Pinet³⁰, R. Pischel¹², F. Poulet⁴, J. Raitala³¹, G. Schwarz³², T. Spohn¹² & S. W. Squyres³³

Affiliations of authors: 1, TU Berlin, D-10623 Germany; 2, Vernadsky Institute-RAS, Moscow 117975, Russia; 3, IFSI/CNR, 00133 Rome, Italy; 4, IAS, 91405 Orsay cedex, France; 5, TU Dresden D-01062, Germany; 6, USGS, MS 975 Menlo Park, California 94025, USA; 7, Universität der Bundeswehr München D-85577, Germany; 8, JPL, Pasadena, California 91109, USA; 9, TU München D-80333, Germany; 10, ESTEC/SCI-SR, Postbus 299, NL-2200 AG Noordwijk, The Netherlands; 11, Arizona State University, Box 871404, Tempe, Arizona 85287-1404, USA; 12, DLR, Berlin D-12489, Germany; 13, Brown University, Box 1846, Providence, Rhode Island 02912, USA; 14, Universität Hannover, D-30167, Germany; 15, Kobe University, 657-8501 Kobe, Japan; 16, National Central University (NCU), Chung-Li 320, Taiwan; 17, IDG-RAS, Moscow 117979, Russia; 18, MPS, Postfach 20, D-37191, Germany; 19, USGS, Flagstaff, Arizona 86001, USA; 20, TU Wien A-1040, Austria; 21, TU Clausthal D-38678, Germany; 22, University of Helsinki, PO Box 14, SF-00014 Finland; 23, Orsay/TERRE, F-91405 Orsay cedex, France; 24, PSI-NW, Winthrop, Washington 98862, USA; 25, University College London WC1E 6BT, UK; 26, The Open University, Milton Keynes MK7 6AA, Buckinghamshire, UK; 27, Universität Köln, D-50923, Germany; 28, FU Berlin D-12249, Germany; 29, IRSPS, I-65127 Pescara, Italy; 30, Observatoire de Midi-Pyrénées, F-31400 Toulouse, France; 31, University of Oulu, FIN-90401, Finland; 32, DLR, D-82234 Wessling, Germany; 33, Cornell University, Ithaca, New York 14853-1301, USA

Current measurement by real-time counting of single electrons

Jonas Bylander, Tim Duty & Per Delsing

Department of Microtechnology and Nanoscience (MC2), Chalmers University of Technology, SE-412 96 Göteborg, Sweden

The fact that electrical current is carried by individual charges has been known for over 100 years, yet this discreteness has not been directly observed so far. Almost all current measurements involve measuring the voltage drop across a resistor, using Ohm's law, in which the discrete nature of charge does not come into play. However, by sending a direct current through a micro-electronic circuit with a chain of islands connected by small tunnel junctions, the individual electrons can be observed one by one. The quantum mechanical tunnelling of single charges in this one-dimensional array is time correlated^{1–3}, and consequently the detected signal has the average frequency $f = I/e$, where I is the current and e is the electron charge. Here we report a direct observation of these time-correlated single-electron tunnelling oscillations, and show electron counting in the range 5 fA–1 pA. This represents a fundamentally new way to measure extremely small currents, without offset or drift. Moreover, our current measurement, which is based on electron counting, is self-calibrated, as the measured frequency is related to the current only by a natural constant.

In the mid-1980s, it was suggested¹ that a small current consisting of individual electrons, tunnelling through a small tunnel junction, could at low temperatures result in an oscillating voltage of amplitude e/C , where C is the capacitance of the tunnel junction. The full theory for these so-called single-electron tunnelling oscillations was then developed², based on earlier work on Bloch oscillations and the underlying Coulomb blockade^{4,5}. This phenomenon of single-electron tunnelling oscillations is similar to the a.c. Josephson effect, as phase and charge are quantum conjugated variables. However, the duality is not complete because the single-electron tunnelling oscillations are lacking coherence. A few years later, these oscillations were detected indirectly by phase locking to an external microwave signal⁶. Shortly thereafter, new devices such as the single-electron turnstile⁷ and the single-electron pump⁸ were invented in order to create a current given by the fundamental relation $I = ef$. Since then, the single-electron pump has been refined to a very high accuracy⁹.

A number of authors have also proposed^{10–13} that it should be possible to turn this relation around, and instead measure the current by monitoring the individual electrons as they pass through a circuit. More recently, single-electron tunnelling events have been observed^{14,15}. In those experiments, however, there was no time correlation, and thus no relation between frequency and current could be demonstrated.

In order to measure current by electron counting, three main ingredients are necessary: time correlation of the tunnelling events, a fast and sensitive charge detector, and a very stable current bias. To bring about time correlation in a single tunnel junction, in contrast to uncorrelated shot noise¹⁶, care must be taken to make the electromagnetic impedance seen by the junction large² compared to the Klitzing resistance, $R_K = h/e^2 \approx 25.8$ k Ω . This can be achieved by placing small-size resistors in close proximity to the junction^{17,18} or by using a one-dimensional series array of tunnel junctions¹⁹.

In our experiment, we have used a superconducting array containing $N = 50$ junctions (Fig. 1). The capacitance of each junction is $C_A \approx 0.42$ fF, and the stray capacitance of an electrode inside the array is $C_0 \approx 30$ aF. In such an array, excess charge on one island polarizes the neighbouring islands, so that the charges repel each

other. The charge is localized, but the potential resulting from the capacitive voltage division in the array is spread over a number of junctions, M . This potential distribution is often referred to as a charge soliton³ (or antisoliton, if a charge is missing), as it can move (by tunnelling) throughout the array without changing its form. In our case, the soliton length is $M \approx \sqrt{C_A/C_0} \approx 3.7$, and the array is in the long limit, $N \gg M$. When the array is biased above a certain threshold voltage, V_b , charges enter from the edge, and as the solitons in the array repel one another, a moving quasi-Wigner lattice of charge solitons is formed. It was shown theoretically^{3,20}, and observed indirectly⁶, that this type of array exhibits time-correlated tunnelling events.

In order to detect the tunnelling charges in real time, we have used an improved version of the conventional single-electron transistor^{2,21,22} (SET) as a charge sensor, namely, the radio-frequency SET²³ (Fig. 1). This is the only device with the required sensitivity and speed: it can detect sub-electron charge changes at frequencies well above 10 MHz. One possible way to do electron counting¹³ is to couple the SET capacitively to one of the electrodes inside the array, and thus observe the potential variations of that electrode as the charge solitons pass. However, we have instead chosen to couple the array directly to the middle electrode (island) of the SET as can be seen in Fig. 1, a configuration known as the resistively coupled^{2,24,25} or array-coupled²⁶ SET. In this way, the full electron charge is detected by the SET.

We have designed the device parameters in order to obtain the desired soliton behaviour and at the same time ensure good performance of the SET. The characteristic time for the soliton dynamics is determined by the RC constant of the tunnel junctions in the array. Monte Carlo simulations^{3,27} show that the narrowest linewidth of the single-electron tunnelling oscillations occurs at

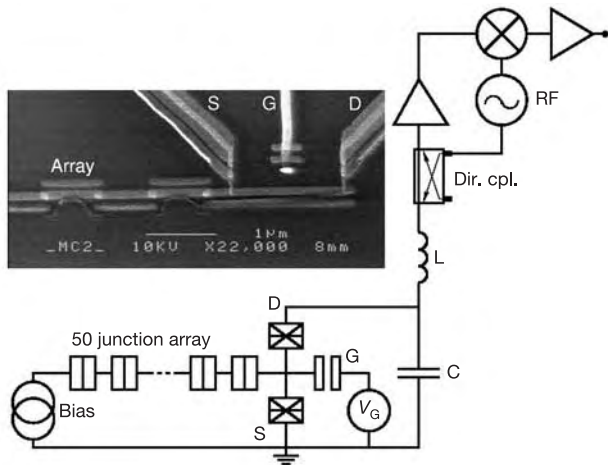


Figure 1 Experimental set-up. Scanning electron micrograph of the sample and a schematic layout of the electron counter based on an SET. Via a directional coupler (dir. cpl.), a radio-frequency signal ($f = 358$ MHz) is applied to the LC circuit (L is inductance) in which the SET is embedded. The quality factor of the resonator is $Q \approx 15$, giving a bandwidth of about 10 MHz. The reflected power is sensitively dependent on the charge on the SET island (the metallic strip connected to the source (S) and drain (D) electrodes of the SET and to the array). For tuning the working point of the SET, a voltage is applied to the capacitive gate (G). The reflected signal is first preamplified, and then mixed with a local oscillator. We measured the small-signal charge sensitivity of the SET to be $2 \times 10^{-5} \text{ eHz}^{-1/2}$ by applying a 200 kHz sine wave of amplitude 0.02 e (r.m.s.) to the gate G (at $B_{||} = 475$ mT). The charging energy of the SET is $E_C/k_B = e^2/2C_{\Sigma}k_B = 1.6$ K, where C_{Σ} is the total capacitance of the island. Either voltage bias (using a transimpedance amplifier (Stanford Research 570) with a d.c. source) or current bias (with a Keithley 263 Calibrator/Source) can be applied to the array of 50 very small tunnel junctions in order to create a current. Injecting the charge directly into the SET island ensures very good coupling to the measurement device. The measurements were performed in a dilution refrigerator at $T \approx 30$ mK. Scale bar in image, $1 \mu\text{m}$.

approximately 1% of the characteristic e/RC current. To keep this current low enough for the SET to follow the oscillations, it is important that the array tunnel junctions have a high resistance. For the SET, on the other hand, sensitivity falls off as the resistance increases. To obtain different oxide thickness for the array and SET junctions we have therefore used a three-angle shadow evaporation process to deposit aluminium²⁸. For the sample described here, the SET source-drain resistance was $R_{\text{SET}} = 30 \text{ k}\Omega$, while the array resistance was $R_N = 0.94 \text{ M}\Omega$ per junction. In Fig. 2, we show the array current-voltage characteristics.

The measurements presented here are done under a parallel magnetic field $B_{||} = 475$ mT, at which the array is still superconducting. This makes the system less sensitive to bias voltage fluctuations, because in the superconducting state the onset of current is rounded, while in the normal state it is quite steep (Fig. 2 inset). Moreover, the sensitivity of the SET is higher in the superconducting state. In general, both electron and Cooper-pair tunnelling can take place in the array, but at this magnetic field the threshold for electron injection is lower than that for Cooper pairs. The suppressed superconducting gap Δ (here $\Delta/k_B \approx 0.6$ K, where k_B is Boltzmann's constant) and the high impedance of the array junctions also suppress the tunnelling rates for Cooper pairs, and we expect to see only electrons at this field.

When the array is biased slightly above the threshold voltage for

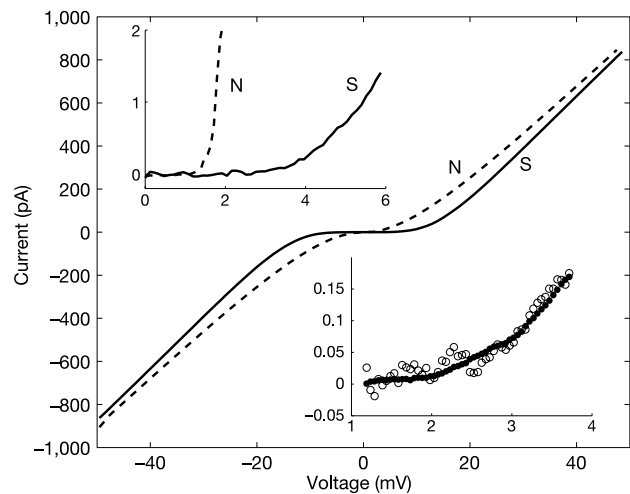


Figure 2 Array current-voltage characteristics in the superconducting state (S: $B_{||} = 475$ mT, solid line) and in the normal state (N: 1.50 T, dashed line). The critical field is $B_{||,c} \approx 650$ mT. The asymptotic normal resistance is $47 \text{ M}\Omega$, and the offset voltage $V_{\text{off}} = 190 \mu\text{V}$ per junction gives the charging energy $E_C^{\text{array}}/k_B = e^2/2C_A k_B = 2.2$ K. Insets have the same units as the main graph. Upper inset; above threshold, the current onset is sharp in the normal state, but more gradual in the superconducting state owing to the subgap resistance R_{SG} , which is both magnetic-field- and voltage-dependent and much higher than the normal resistance R_N . Consequently, the sensitivity to bias voltage fluctuations is much higher in the normal state. Ignoring the effect of background charges, the threshold voltage for electrons is theoretically $V_t^e = Kme/2C_A + (1 + M)\Delta(B)/e \approx 0.87$ mV, where $K = (\sqrt{1 + 1/4M^2} + 1/2M)^{-1} \approx 0.9$, whereas for Cooper pairs it is $V_t^{2e} = Kme/C_A \approx 1.2$ mV. The effect of random background charges further complicates the situation by altering these thresholds^{30,31}. Owing to the weak Josephson coupling energy, $E_J/k_B = 1 \text{ mK} \ll T$, and the high impedance of the neighbouring junctions (we assume a subgap resistance $R_{SG} = 100R_N$), the rate for Cooper pair tunnelling is much lower than the rate for electron tunnelling above V_t^{2e} . Lower inset; comparison of d.c. current measurement by counting individual electrons, that is, measuring a frequency (filled circles), and a conventional d.c. current measurement (open circles) by voltage biasing through a Stanford Research 570 transimpedance amplifier. Note that for very low currents, the spread in the frequency measurements is smaller than in the conventional current measurements.

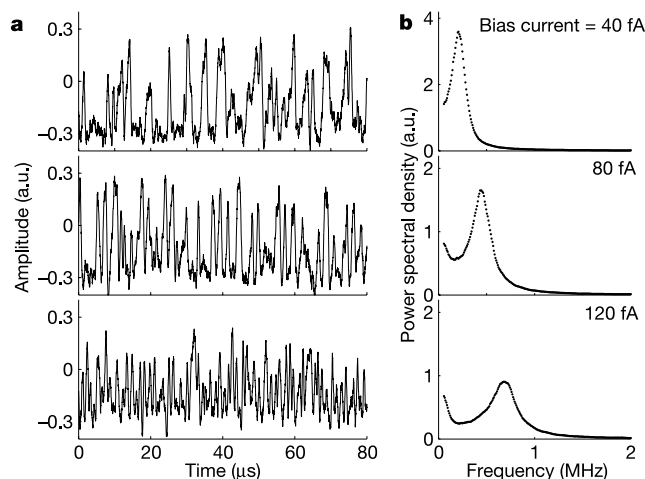


Figure 3 Experimental data. **a**, Real-time electron counting. Each peak in the reflected signal corresponds to one electron tunnelling into the SET island. The traces are snapshots recorded on a digital oscilloscope. The signal is low-pass filtered with a cut-off frequency of 15 MHz and then digitally smoothed with a moving average of 0.5 μ s. The applied currents were, from top to bottom, 40, 80 and 120 fA, corresponding to the frequency ($f = I/e$) 250, 500 and 750 kHz, respectively. **b**, Power spectral densities of the reflected power from the SET, measured using a spectrum analyser with 100 kHz resolution bandwidth. The peak frequencies are $f = 236, 460$ and 700 kHz, corresponding to the currents 38, 74 and 112 fA, respectively. The deviation from the nominal bias currents is due to a small offset in the current source. a.u., arbitrary units.

electrons, V_c^* , a small (sub-pA) current starts to flow. As the charges approach the SET, the source-drain current is modulated by the charge induced on the middle island, resulting in modulation of the reflected radio-frequency power. In this way, we have detected single-electron tunnelling oscillations in real time. Figure 3a shows time traces of single-electron tunnelling events for three different currents. In the range from 5 fA to 1 pA, the reflected power spectrum shows a peak (Fig. 3b) at a frequency that corresponds to the applied current in the array, $f = I/e$ (Fig. 4). This demonstrates time correlation between consecutive tunnelling events. The signal is gradually smeared by temperature, but persists even at 300 mK. The peak broadens considerably when the current approaches 10% of the e/RC current, in fair agreement with simulations, assuming a subgap resistance $R_{SG} = 100R_N = 94$ M Ω .

Our preamplifier, with a noise temperature of 2 K, sets the overall noise level in our measurement system. The counting errors caused by the limited sensitivity of the SET can be estimated¹³ as $P = 0.5\text{erfc}(q_c/Q_N\sqrt{2})$, where Q_N is the charge noise within the measurement bandwidth, q_c is the charge threshold for a count to be registered, and erfc is the complementary error function. For realistic parameters, this expression gives an accuracy of 1 part in 10^6 . However, P is very strongly dependent on the sensitivity of the SET, which suggests that metrological accuracy can be achieved in an optimized device. In this experiment, there are obviously other error sources limiting the accuracy, such as bias instability that smears the peak in the spectrum, and background charges that affect both the operation point of the SET and the soliton flow in the array.

An important property of the electron counter described here is its dynamic current range. At low currents (<3 fA, according to our simulations²⁷), when there is only one soliton in the array at a time, space correlation—and hence also time correlation—is lost. The opposite limit is ultimately set by the e/RC current. In a realistic experiment, however, the resolution bandwidth of our measurement set-up, $1/f$ noise and instability of the bias at very low currents prevent us from measuring with any accuracy below 5 fA. At high currents, the measurement speed is restricted by the bandwidth of the SET.

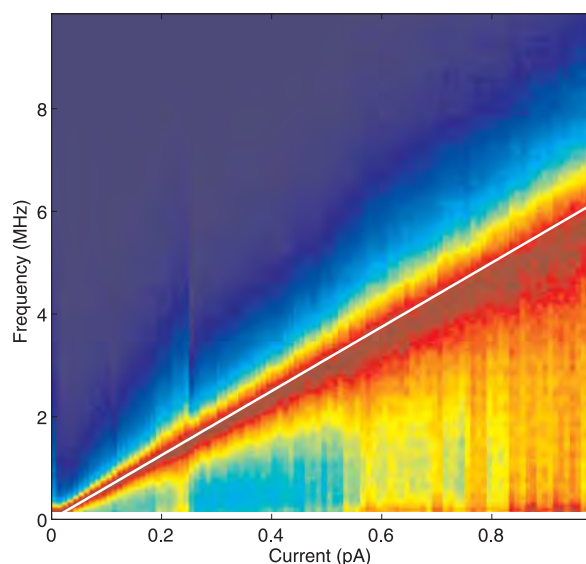


Figure 4 Single-electron tunnelling oscillations. The power spectra (Fig. 3b) have been normalized so that the peaks all have unity height. The dynamic colour range goes from blue to red, corresponding to zero and one, respectively. The position of the red ridge agrees well with the solid white line $f = I/e$, where I is the applied bias current. This demonstrates current measurement by electron counting from 5 fA to 1 pA. The noticeable broadening and increasing asymmetry of the spectral peak for higher currents gives a small discrepancy between the measured data and I/e .

From an applications point of view it would be desirable to measure larger currents, and to decrease the input impedance of the ammeter. This is possible with an electron counter, as it is straightforward to put several devices in parallel, which is not the case with single-electron pumps. Then it could also be used to close the quantum metrological triangle⁵ relating current, voltage and frequency by fundamental constants. Another important feature is that the counter is self-calibrated, and does not suffer from problems like offset or drift in the measurement set-up.

Recent interest in quantum noise in mesoscopic systems has led to the theory of full counting statistics²⁹, in which general information about charge transfer can be gained from the higher statistical moments of noise. So far, very few experiments have been performed along these lines, but the relatively low bandwidth of the tunnelling process in the array suggests the possibility of measuring the full statistics of the charge transport, and gaining access to higher statistical moments of the current. $\square$

Received 11 November 2004; accepted 20 January 2005; doi:10.1038/nature03375.

- Ben-Jacob, E. & Gefen, Y. New quantum oscillations in current driven small junctions. *Phys. Lett. A* **108**, 289–292 (1985).
- Averin, D. V. & Likharev, K. K. Coulomb blockade of single-electron tunneling, and coherent oscillations in small tunnel junctions. *J. Low Temp. Phys.* **62**, 345–373 (1986).
- Likharev, K. K., Bakhalov, N. S., Kazach, G. S. & Serdyukova, S. I. Single electron tunnel junction array: An electrostatic analog of the Josephson transmission line. *IEEE Trans. Magn.* **25**, 1436–1439 (1989).
- Averin, D. V., Zorin, A. B. & Likharev, K. K. Bloch oscillations in small Josephson junctions. *Sov. Phys. JETP* **61**, 407–413 (1985).
- Likharev, K. K. & Zorin, A. B. Theory of the Bloch-wave oscillations in small Josephson junctions. *J. Low Temp. Phys.* **59**, 347–382 (1985).
- Delsing, P., Likharev, K. K., Kuzmin, L. S. & Claeson, T. Time correlated single electron tunneling in one-dimensional arrays of ultrasmall tunnel junctions. *Phys. Rev. Lett.* **63**, 1861–1864 (1989).
- Geerlings, L. J. et al. Frequency-locked turnstile device for single electrons. *Phys. Rev. Lett.* **64**, 2691–2694 (1990).
- Pothier, H., Lafarge, P., Urbina, C., Estève, D. & Devoret, M. H. Single-electron pump based on charging effects. *Europhys. Lett.* **17**, 249–254 (1992).
- Keller, M. W., Martinis, J. M., Zimmerman, N. M. & Steinbach, A. H. Accuracy of electron counting using a 7-junction electron pump. *Appl. Phys. Lett.* **69**, 1804–1806 (1996).
- Likharev, K. K. in *Granular Nanoelectronics* (eds Ferry, D., Barker, J. R. & Jacoboni, C.) 371–391 (Plenum, New York, 1991).
- Flensberg, K., Odintsov, A. A., Liefrink, E. & Teunissen, P. Towards single-electron metrology. *Int. J. Mod. Phys. B* **13**, 2651–2687 (1999).

12. Visscher, E. H. *et al.* Broadband single-electron tunneling transistor. *Appl. Phys. Lett.* **68**, 2014–2016 (1996).
13. Visscher, E. *Technology and Applications of Single-Electron Tunneling Devices*. Ph.D thesis, Delft Univ. Technol. (1996).
14. Lu, W., Ji, Z. Q., Pfeiffer, L., West, K. W. & Rimberg, A. J. Real-time detection of electron tunnelling in a quantum dot. *Nature* **423**, 422–425 (2003).
15. Fujisawa, T., Hayashi, T., Hirayama, Y., Cheong, H. D. & Jeong, Y. H. Electron counting of single-electron tunneling current. *Appl. Phys. Lett.* **84**, 2343–2345 (2004).
16. Blanter, Ya. M. & Büttiker, M. Shot noise in mesoscopic conductors. *Phys. Rep.* **336**, 1–166 (2000).
17. Cleland, A. N., Schmidt, J. M. & Clarke, J. Charge fluctuations in small-capacitance junctions. *Phys. Rev. Lett.* **64**, 1565–1568 (1990).
18. Kuzmin, L. S., Nazarov, Yu. V., Haviland, D. B., Delsing, P. & Claeson, T. Coulomb blockade and incoherent tunneling of Cooper pairs in ultrasmall junctions affected by strong quantum fluctuations. *Phys. Rev. Lett.* **67**, 1161–1164 (1991).
19. Delsing, P., Likharev, K. K., Kuzmin, L. S. & Claeson, T. Effect of high-frequency electrodynamic environment on the single-electron tunneling in ultrasmall junctions. *Phys. Rev. Lett.* **63**, 1180–1183 (1989).
20. Korotkov, A. N. Analytical calculation of single-electron oscillations in one-dimensional arrays of tunnel junctions. *Phys. Rev. B* **50**, 17674–17677 (1994).
21. Likharev, K. K. Single-electron transistors: Electrostatic analogs of the dc-squids. *IEEE Trans. Magn.* **23**, 1142–1145 (1987).
22. Fulton, T. A. & Dolan, G. J. Observation of single-electron charging effects in small tunnel junctions. *Phys. Rev. Lett.* **59**, 109–112 (1987).
23. Schoelkopf, R. J., Wahlgren, P., Kozhevnikov, A. A., Delsing, P. & Prober, D. E. The radio-frequency single-electron transistor (RF-SET): A fast and ultrasensitive electrometer. *Science* **280**, 1238–1242 (1998).
24. Pashkin, Y. A., Nakamura, Y. & Tsai, J. S. Metallic resistively coupled single-electron transistor. *Appl. Phys. Lett.* **74**, 132–134 (1999).
25. Korotkov, A. N. Theoretical analysis of the resistively coupled single-electron transistor. *Appl. Phys. Lett.* **72**, 3226–3228 (1998).
26. Delsing, P., Claeson, T., Kazach, G. S., Kuzmin, L. S. & Likharev, K. K. 1-D array implementation of the resistively-coupled single-electron transistor. *IEEE Trans. Magn.* **27**, 2581–2584 (1991).
27. Bakhvalov, N. S., Kazach, G. S., Likharev, K. K. & Serdyukova, S. I. Single-electron solitons in one-dimensional tunnel structures. *Sov. Phys. JETP* **68**, 581–587 (1989).
28. Dolan, G. J. Offset masks for lift-off photoprocessing. *Appl. Phys. Lett.* **31**, 337–339 (1977).
29. Levitov, L. S., Lee, H. & Lesovik, G. B. Electron counting statistics and coherent states of electric current. *J. Math. Phys.* **37**, 4845–4866 (1996).
30. Middleton, A. A. & Wingreen, N. S. Collective transport in arrays of small metallic dots. *Phys. Rev. Lett.* **71**, 3198–3201 (1993).
31. Kaplan, D. M., Sverdlov, V. A. & Likharev, K. K. Coulomb gap, Coulomb blockade, and dynamic activation energy in frustrated single-electron arrays. *Phys. Rev. B* **68**, 45321 (2003).

Acknowledgements We thank C. Kristofferson, S. Pedersen and P. Wahlgren for assistance in the early stages of this work, K. Bladh, D. Gunnarsson, S. Kafanov and M. Taslakov for technical assistance, and T. Claeson, H. Nilsson, K.-E. Rydler, G. Wendin, C. Wilson and A. Zorin for discussions. The work was supported by the Swedish SSF and VR, by the EU research project COUNT and by the Wallenberg foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.B. (jonas.bylander@mc2.chalmers.se).

Relaxor ferroelectricity and colossal magnetocapacitive coupling in ferromagnetic CdCr_2S_4

J. Hemberger¹, P. Lunkenheimer¹, R. Fichtl¹, H.-A. Krug von Nidda¹, V. Tsurkan^{1,2} & A. Loidl¹

¹Experimental Physics V, Centre for Electronic Correlations and Magnetism, University of Augsburg, 86159 Augsburg, Germany

²Institute for Applied Physics, Academy of Sciences of Moldova, MD-2028, Chisinau, Moldova

Materials in which magnetic and electric order coexist—termed ‘multiferroics’ or ‘magnetoelectrics’—have recently become the focus of much research^{1–4}. In particular, the simultaneous occurrence of ferromagnetism and ferroelectricity, combined with an intimate coupling of magnetization and polarization via magnetocapacitive effects, holds promise for new generations

of electronic devices. Here we present measurements on a simple cubic spinel compound with unusual, and potentially useful, magnetic and electric properties: it shows ferromagnetic order coexisting with relaxor ferroelectricity (a ferroelectric cluster state with a smeared-out phase transition), both having sizable ordering temperatures and moments. Close to the ferromagnetic ordering temperature, the magnetocapacitive coupling (characterized by a variation of the dielectric constant in an external magnetic field) reaches colossal values, approaching 500 per cent. We attribute the relaxor properties to geometric frustration, which is well known for magnetic moments but here is found to impede long-range order of the structural degrees of freedom that drive the formation of the ferroelectric state.

Preparation of a material in which ferroelectricity and ferromagnetism coexist would be a milestone for modern electronics and functionalized materials. The most appealing applications of such a material are new types of storage media using both magnetic and electric polarization, and the possibility of electrically reading/writing magnetic memory devices (and vice versa). However, it is clear now that ferroelectric ferromagnets are rare^{5,6} and mostly exhibit rather weak ferromagnetism. Spinel compounds are an important class of materials, and their electronic properties have been the focus of research since the work of Verwey on magnetite⁷. Recent reports of geometrical frustration of the spin and orbital degrees of freedom in sulpho-spinels⁸, and the observation of an orbital-glass state⁹ in these compounds, demonstrate the rich and complex physics that is characteristic of this system. Here we report another interesting experimental observation in a spinel system: relaxor ferroelectricity in ferromagnetic CdCr_2S_4 and the occurrence of colossal magnetocapacitive effects.

CdCr_2S_4 crystallizes in the normal cubic spinel structure (space group $Fd\bar{3}m$, lattice constant $a = 1.024$ nm), with Cr^{3+} octahedrally surrounded by sulphur ions, yielding a half-filled lower t_{2g} triplet with spin $S = 3/2$. Ferromagnetism in CdCr_2S_4 is well known¹⁰, but early experimental observations of a number of unusual features seem to have been overlooked: for example, reports of an anomalous expansion coefficient at low temperatures^{11,12}, an unexpected concomitant broadening of the diffraction lines¹¹, a strong blue shift of the absorption edge on passing the ferromagnetic phase transition¹³, the observation of anomalously large phonon shifts and damping effects close to the magnetic transition temperature T_c (ref. 14), and the observation of large magnetoresistance effects¹⁵.

Figure 1a shows the inverse magnetic susceptibility χ^{-1} and the low-temperature (low- T) magnetization M . The straight line indicates a fit to the paramagnetic susceptibility, which results in a Curie–Weiss temperature of 155 K and a paramagnetic moment of $3.88 \mu_B$, close to the theoretically expected value of $3.87 \mu_B$ per Cr^{3+} . The deviations when approaching T_c arise from ferromagnetic fluctuations. $M(T)$ strongly increases at T_c , the further increase just below T_c being interrupted by demagnetization effects. The inset in Fig. 1a shows the ferromagnetic hysteresis, revealing soft-magnetic behaviour and a saturation magnetization of $3 \mu_B$ per Cr^{3+} . Hence, CdCr_2S_4 behaves like a typical soft ferromagnet.

Figure 1b shows the temperature dependence of the dielectric constant ϵ' , measured at 9.5 Hz (right scale). $\epsilon'(T)$ exhibits a steep rise below T_c , obviously driven by the onset of ferromagnetic order. Strong magnetoelectric coupling effects must be responsible for this behaviour. At lower temperatures, $\epsilon'(T)$ decreases again, leading to a maximum reminiscent of ferroelectric behaviour. However, its rather large width and the only moderately high maximum values of $\epsilon' \approx 100$ indicate a transition into a polar state of diffusive character. Indeed, a finite polarization, a direct proof of ferroelectricity, arises well below T_c (Fig. 1b, left scale). Final evidence is provided in Fig. 1c, which shows a polarization cycle (using electric fields up to 400 kV m^{-1}) exhibiting a clear ferroelectric hysteresis. But the relatively slim loop indicates non-canonical ferroelectric behaviour¹⁶.

Evidence for the actual character of this ferroelectric state is revealed by the temperature dependence of ϵ' , shown in Fig. 2a for a series of measuring frequencies. A very broad peak appears between 150 and 250 K, shifting to lower temperatures and increasing in amplitude with decreasing frequency. It exhibits the characteristic behaviour of a relaxor ferroelectric, the typical strong dispersion effects often being ascribed to the freezing-in of ferroelectric clusters^{16,17}. In contrast to conventional ferroelectrics, relaxor ferroelectrics are characterized by a diffuse phase transition (extending over a finite temperature range), by strong relaxational dispersion effects in dielectric constant and loss, and by a macroscopic polarization showing up only well below the transition temperature^{16–18}. Most relaxor ferroelectrics known so far are perovskite-related materials where long-range polar order is suppressed by substitutional disorder¹⁷. However, the relaxor properties in CdCr_2S_4 occur in a pure compound without any disorder. By comparison to measurements with silver-paint contacts (not shown), the increase of ϵ' at temperatures above those at which the peak occurs is found to be due to contact effects¹⁹, whereas the peak itself is clearly an intrinsic property of CdCr_2S_4 . The high-temperature flank of the relaxor peaks can be roughly described by a Curie–Weiss law with a characteristic temperature of 135 K (dashed line in Fig. 2a). In relaxors, deviations from a Curie–Weiss law are often observed; but this law gives a rough estimate of a quasistatic freezing temperature. Notably, significant spontaneous polarization sets in only well below 135 K, a behaviour often found in relaxor ferroelectrics¹⁷.

Further evidence for typical relaxor behaviour arises from the frequency (ν) dependence of the dielectric loss ϵ'' . The inset of Fig. 2a shows $\epsilon''(\nu)$ for a series of temperatures, obtained after subtraction of a conductivity contribution. Broad peak maxima are revealed, typical of relaxational behaviour. The peak frequency decreases when the temperature is lowered, indicating the freezing-in of polar moments. In CdCr_2S_4 , the half-width of the loss peak at 151 K amounts to nearly three decades in frequency, significantly broader than the 1.14 decades expected for a Debye relaxation, indicating a broad distribution of relaxation times. However, in most canonical relaxors even broader distributions are found^{17,18}, and we cannot exclude the possibility that contributions from

hopping charge carriers (for example, small polarons trapped in defects²⁰) play a role in the appearance of the observed relaxation feature.

A second relaxation-derived dispersion with a concomitant strong increase of ϵ' appears at temperatures below the onset of ferromagnetic order (Fig. 2a). A double peak appearing at low frequencies becomes increasingly suppressed at higher frequencies. This can be explained by assuming that the polar moments, which at about 100 K are frozen-in at the timescale of the experiment, speed up again below T_c . The increasing ferromagnetic correlations accelerate the mean polar relaxation rates. This leads to a remelting of the frozen-in relaxor state at $T < T_c$, resulting in strongly enhanced dielectric constants. Only when the temperature is further lowered (below about 20 K) does temperature finally 'win', and ϵ' becomes reduced owing to a complete freezing-in of the polar dynamics. Most probably, exchangestriction (volume changes arising from the magnetic exchange energy) softens the lattice, reduces the energy barriers against dipolar reorientation and enhances the mean relaxation rates. Exchangestriction was also used to explain the enormous blue shift in CdCr_2S_4 (ref. 21). On the other hand, a contribution from hopping-type charge transport to the anomaly at T_c by way of a variation of the a.c. conductivity, also affecting ϵ' , cannot be excluded, especially as it is well known that CdCr_2S_4 exhibits sizable magnetoresistance effects¹⁵. In any case, overall the magnetocapacitive coupling in CdCr_2S_4 is different to all mechanisms observed so far and purely dynamic in origin.

Figure 2b demonstrates that the magnetocapacitive coupling is indeed extremely strong and can be termed colossal. It shows the temperature dependence of the dielectric constant at 9.5 Hz and 17 kHz as measured in zero magnetic field and in a field of 5 T. The relaxation dynamics at $T > T_c$ remain essentially unaffected by the field. However, strong magnetocapacitive effects are observed close to T_c . The magnetocapacitance—as derived from the difference of the dielectric constant in zero and non-zero fields—is shown in the inset of Fig. 2b. At low measuring frequencies, the maximum effects occurring close to T_c reach 450% at 9.5 Hz. For possible applications of this phenomenon, some hurdles certainly have to be overcome: for example, by doping or searching for related multi-ferroic spinel systems with higher transition temperatures.

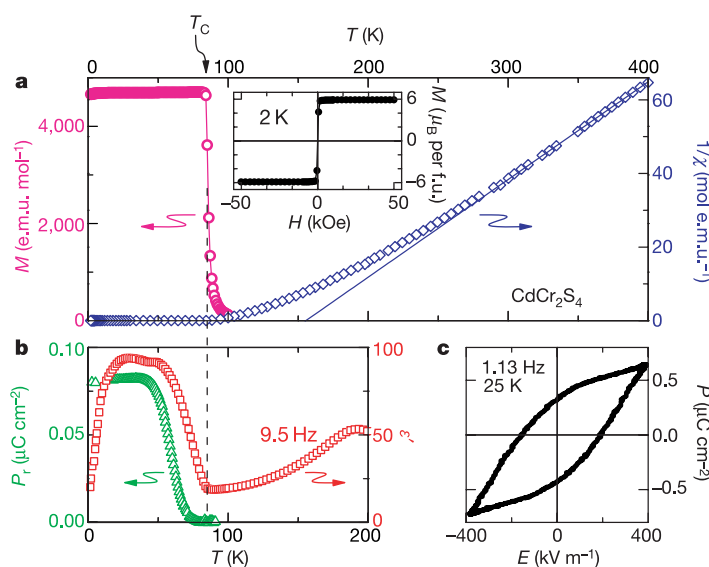


Figure 1 Magnetic and dielectric characterization of CdCr_2S_4 . **a**, Right scale: inverse magnetic susceptibility versus temperature measured at 100 Oe. The solid line indicates a fit ($T > 300$ K) using a Curie–Weiss law, with a Curie–Weiss temperature of 155 K. Left scale: magnetization versus temperature measured at 100 Oe. Inset: ferromagnetic hysteresis at 2 K, indicating a saturated moment of $6 \mu_B$ per formula unit (f.u.). **b**, Right

scale: dielectric constant versus temperature measured at a frequency of 9.5 Hz. Left scale: thermo-remnant polarization versus temperature measured after cooling in an electric field of 50 kV m^{-1} . **c**, Polarization versus electric field, showing a ferroelectric hysteresis.

Finally, Fig. 3 shows the thermo-remnant polarization (P) for various magnetic fields, demonstrating a strong coupling of both quantities. Increasing external magnetic fields suppress the polar state, in marked contrast to the findings in other magnetocapacitive materials^{1–3}. This is fully consistent with our assumption that the onset of ferromagnetic order enhances the reorientational mobility of the polar entities. Increasing magnetic fields drive the system towards ferromagnetism, and thus the macroscopic polarization breaks down at lower temperatures for higher fields (Fig. 3). The polarization shift increases with magnetic field (inset of Fig. 3), showing the strong competition between ferroelectricity and ferromagnetism. The magnetic-field dependence shown in Figs 2b and 3 was measured with the electric field perpendicular to the magnetic field; measurements with the electric field parallel to the magnetic field led to identical results.

Two more questions remain to be resolved: what is the nature of the ordered dipolar moment in CdCr_2S_4 , and how could relaxor ferroelectricity occur without disorder-derived frustration? It is clear that the polar ordering must be displacive in nature, driven by soft-mode behaviour as in ferroelectric perovskites. This notion is consistent with the above-mentioned structural softening deduced in ref. 11. Indeed, the observation of polar displacements of the octahedrally coordinated ions has been reported for a number

of spinels²². Guided by these results, we suggest that ferroelectricity in CdCr_2S_4 results from an off-centre position of the Cr^{3+} ions. It may well be that spinel compounds will turn out to be a new class of semiconducting ferroelectrics.

In relaxor ferroelectrics, the freezing of polar moments is driven by frustrated interactions related to disorder. For magnetic systems, frustration can also arise in ordered solids from geometrical constraints alone²³. Ising spins with nearest-neighbour antiferromagnetic interactions on a triangular lattice are a standard example. The possibility that geometric frustration might also exist in non-magnetic systems has been suggested²⁴, exemplified by ZrW_2O_8 revealing negative thermal expansion²⁵. This effect has been explained with a concept of frustrated soft-mode behaviour²⁶. CdCr_2S_4 also shows negative thermal expansion, and reveals a strong broadening of diffraction lines below 120 K (refs 11,12). The broadening of diffraction lines on cooling provides clear evidence for the loss of true long-range order, and indicates that local order may differ substantially from global symmetry. Similar effects have been observed²⁷ in the orientational glass KBr:KCN . In orientational glasses²⁸, disorder-derived frustration drives the glassy freezing. In contrast, the suggested polar freezing in CdCr_2S_4 could be governed by geometrical constraints.

The most interesting aspects of the behaviour of CdCr_2S_4 are the coexistence of ferromagnetism and proper ferroelectricity (with reasonable ordering temperatures and sizable saturated moments), colossal magnetocapacitive effects, and strong coupling of ferroelectric polarization to external magnetic fields. Compared to recent reports on other multiferroics^{1–4}, this spinel system is unique. The perovskite-derived manganites treated in refs 1–3 are improper ferroelectrics with polar order induced by modulated spin structures, accompanied by lattice modulations. In contrast, the hexagonal manganite of ref. 4 is a proper ferroelectric with high ordering temperature ($T_c = 875$ K) and much lower antiferromagnetic ordering ($T_N = 75$ K). Ferromagnetism is induced by applied electric fields. In CdCr_2S_4 , ferromagnetism and ferroelectricity to first order develop independently, driven by electronic superexchange and soft-mode behaviour, respectively. Both phenomena are linked to the Cr^{3+} ions and strongly compete, yielding colossal magnetoelectric effects. In CdCr_2S_4 , magnetization and polarization, as well as dipolar and magnetic susceptibilities, are compatible with a global cubic symmetry. All quantities are independent of the

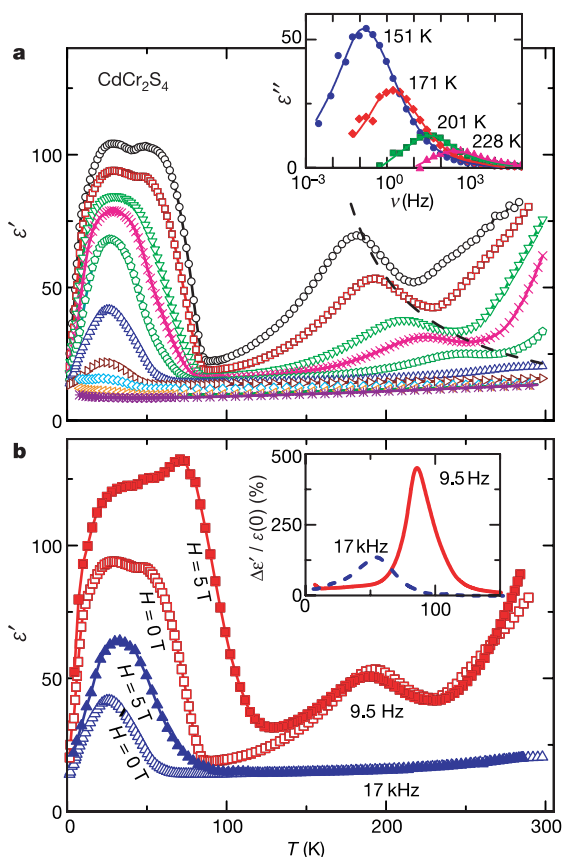


Figure 2 Magnetocapacitive behaviour of CdCr_2S_4 . **a**, Temperature dependence of ϵ' at various frequencies (from top to bottom: 3 Hz, 9.5 Hz, 53 Hz, 170 Hz, 950 Hz, 17 kHz, 950 kHz, 12 MHz, 83 MHz, 3 GHz). The dashed line indicates the static dielectric susceptibility following a Curie–Weiss-like law with a characteristic temperature of 135 K. The inset shows $\epsilon''(\nu)$ for various temperatures above 150 K (lines are to guide the eye). **b**, Dielectric constant versus temperature at 9.5 Hz and 17 kHz, measured at zero field and in an external magnetic field of 5 T, directed perpendicular to the electric field. The inset provides a measure of the magnetocapacitive effects, with $\Delta\epsilon' = \epsilon'(5\text{T}) - \epsilon'(0\text{T})$, for the two frequencies shown in the main panel.

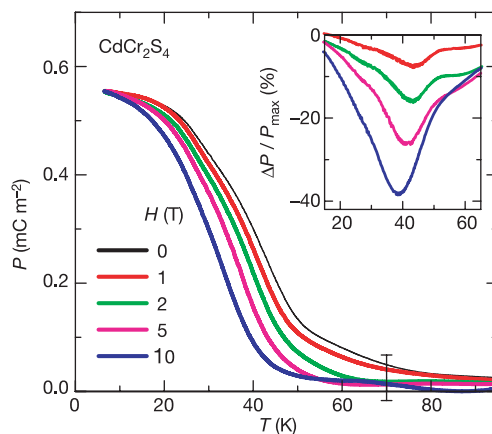


Figure 3 Thermo-remnant polarization versus temperature. Measurements were made after cooling the sample in an electric field of 150 kV m^{-1} in various external magnetic fields, directed perpendicular to the electric field. The error bar indicates an 80% confidence error caused by conductivity contributions. The inset shows the magnetic-field-dependent variation of the polarization, $\Delta P = P(H) - P(0\text{T})$, normalized to the maximum polarization of about 0.55 mC m^{-2} . Colour coding as in main panel; x axis shows temperature in K.

direction of the magnetic field with respect to the electric field, indicating that CdCr_2S_4 belongs to a new class of multiferroics. Despite these differences, frustration seems to play a fundamental role in all these multiferroic compounds: it is frustration of the electronic degrees of freedom in the manganites, and geometrical frustration of the lattice degrees of freedom in the spinel. □

Methods

Single crystals of CdCr_2S_4 were grown by chemical transport reaction, using chlorine as the transport agent. The starting material was the ternary compound obtained by standard ceramic techniques. The single crystals were routinely characterized by powder and single-crystal X-ray diffraction, and by wavelength sensitive electron-probe microanalysis. We found an almost ideal stoichiometry, and we can exclude any site inversion between the cations. To check for the robustness of the results, we performed measurements on several crystals from different batches. All samples revealed ferromagnetic and ferroelectric ordering temperatures well within the experimental uncertainties. Magnetic susceptibility and magnetization were measured using a commercial SQUID. For the dielectric measurements of the present work, sputtered gold contacts were applied to the plate-like samples, forming a parallel-plate capacitor. To check for possible contributions from electrode polarization, the same sample was also measured with silver-paint contacts. The dielectric constant and loss were measured over a broad frequency range of nine decades ($3\text{ Hz} < \nu < 3\text{ GHz}$). A frequency-response analyser (Novocontrol α -analyser) was used for frequencies $\nu < 1\text{ MHz}$, and a reflectometric technique using an impedance analyser (Agilent E4291A) for $\nu > 1\text{ MHz}$ (ref. 29). The temperature-dependent polarization was determined via the pyro charge measured with an electrometer (Keithley 617). The same device was used as impedance transformer within a Sawyer-Tower circuit, read out by a two-channel digital oscilloscope (Tektronik TDS210) and stimulated via a high-voltage amplifier (Trek 609). For cooling to temperatures down to 1.4 K , a conventional ^4He bath-cryostat (Cryovac Konti-IT) was used; additional magnetic-field-dependent measurements up to 10 T were performed in an Oxford cryomagnet. The magnetic anisotropy was checked by angular-dependent SQUID and electron-spin-resonance measurements. We found almost isotropic behaviour in the paramagnetic phase, and a slight cubic anisotropy in the magnetically ordered phase. All dielectric measurements were performed with $\mathbf{E} \parallel \mathbf{H}$ and $\mathbf{E} \perp \mathbf{H}$, and we also measured along different crystallographic directions, revealing identical results within experimental resolution.

Received 19 November; accepted 23 December 2004; doi:10.1038/nature03348.

- Kimura, T. *et al.* Magnetic control of ferroelectric polarization. *Nature* **426**, 55–58 (2003).
- Goto, T., Kimura, T., Lawes, G., Ramirez, A. P. & Tokura, Y. Ferroelectricity and giant magnetocapacitance in perovskite rare-earth manganites. *Phys. Rev. Lett.* **92**, 257201 (2004).
- Hur, N. *et al.* Electric polarization in a multiferroic material induced by magnetic fields. *Nature* **429**, 392–395 (2004).
- Lottermoser, Th. *et al.* Magnetic phase control by an electric field. *Nature* **430**, 541–544 (2004).
- Hill, N. A. Why are there so few magnetic ferroelectrics? *J. Chem. Phys. B* **104**, 6694–6709 (2000).
- Smolenskii, G. A. & Chupis, I. E. Ferroelectromagnets. *Sov. Phys. Usp.* **25**, 475–493 (1983).
- Verwey, J. E. W. Electronic conduction of magnetite (Fe_3O_4) and its transition point at low temperature. *Nature* **144**, 327–328 (1939).
- Fritsch, V. *et al.* Spin and orbital frustration in MnSc_2S_4 and FeSc_2S_4 . *Phys. Rev. Lett.* **92**, 116401 (2004).
- Fichtl, R., Tsurkan, V., Lunkenheimer, P., Hemberger, J., Fritsch, V., Krug von Nidda, H.-A., Scheidt, E.-W. & Loidl, A. Orbital freezing and orbital glass state in FeCr_2S_4 . *Phys. Rev. Lett.* **94**, 027601 (2005).
- Baltzer, P. K., Lehmann, H. W. & Robbins, M. Insulating ferromagnetic spinels. *Phys. Rev. Lett.* **15**, 493–495 (1965).
- Göbel, H. Local lattice distortions in chromium chalcogenide spinels at low temperatures. *J. Magn. Mater.* **3**, 143–146 (1976).
- Martin, G. W., Kellogg, A. T., White, R. L. & White, R. M. Exchangestriction in CdCr_2S_4 and CdCr_2Se_4 . *J. Appl. Phys.* **40**, 1015–1016 (1969).
- Harbeke, G. & Pinch, H. Magnetoabsorption in single-crystal semiconducting ferromagnetic spinels. *Phys. Rev. Lett.* **17**, 1090–1092 (1966).
- Wakamura, K. & Arai, T. Effect of magnetic ordering on phonon parameters for infrared active modes in ferromagnetic spinel CdCr_2S_4 . *J. Appl. Phys.* **63**, 5824–5829 (1988).
- Lehmann, H. W. & Robbins, M. Electrical transport properties of the insulating ferromagnetic spinels CdCr_2S_4 and CdCr_2Se_4 . *J. Appl. Phys.* **37**, 1389–1390 (1966).
- Samara, G. A. The relaxation properties of compositionally disordered ABO_3 perovskites. *J. Phys. Condens. Matter* **15**, R367–R411 (2003).
- Cross, L. E. Relaxor ferroelectrics. *Ferroelectrics* **76**, 241–267 (1987).
- Kamba, S. *et al.* Dielectric dispersion of relaxor PLZT ceramics in the frequency range 20 Hz–100 THz. *J. Phys. Condens. Matter* **12**, 497–519 (2000).
- Lunkenheimer, P. *et al.* Origin of apparent colossal dielectric constants. *Phys. Rev. B* **66**, 052105 (2002).
- Austin, I. G. & Mott, N. F. Polarons in crystalline and non-crystalline materials. *Adv. Phys.* **18**, 41–103 (1969).
- Callen, E. Optical absorption edge of magnetic semiconductors. *Phys. Rev. Lett.* **20**, 1045–1048 (1968).
- Grimes, N. W. Off-centre ions in compounds with spinel structure. *Phil. Mag.* **26**, 1217–1226 (1972).
- Ramirez, A. P. in *Handbook of Magnetic Materials* Vol. 13 (ed. Buschow, K. H. J.) 423–520 (Elsevier, Amsterdam, 2001).
- Ramirez, A. P. Magic moments. *Nature* **421**, 483 (2003).
- Mary, T., Evans, J. S. O., Vogt, T. & Sleight, A. W. Negative thermal expansion from 0.3 to 1050 K in ZrW_2O_8 . *Science* **272**, 90–92 (1996).
- Cao, D., Bridges, F., Kowach, G. R. & Ramirez, A. P. Frustrated soft modes and negative thermal expansion in ZrW_2O_8 . *Phys. Rev. Lett.* **89**, 215902 (2002).
- Knorr, K. & Loidl, A. Anomalous diffraction profiles of alkali-halide-alkali-cyanide mixed crystals. *Phys. Rev. Lett.* **57**, 460–462 (1986).

- Höchli, U. T., Knorr, K. & Loidl, A. Orientational glasses. *Adv. Phys.* **39**, 405–615 (1990).
- Lunkenheimer, P., Schneider, U., Brand, R. & Loidl, A. Glassy dynamics. *Contemp. Phys.* **41**, 15–36 (2000).

Acknowledgements This work was supported partly by the Deutsche Forschungsgemeinschaft via the Sonderforschungsbereich 484 and partly by the BMBF via VDI/EKM.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.L. (peter.lunkenheimer@physik.uni-augsburg.de).

Two-electron dissociation of single molecules by atomic manipulation at room temperature

P. A. Sloan* & R. E. Palmer

Nanoscale Physics Research Laboratory, School of Physics and Astronomy, The University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

* Present address: Department of Chemistry, University of Toronto, 80 St George Street, Toronto, Ontario, M5S 3H6, Canada

Using the tip of a scanning tunnelling microscope (STM) to mechanically manipulate individual atoms and molecules on a surface is now a well established procedure^{1,2}. Similarly, selective vibrational excitation of adsorbed molecules with an STM tip to induce motion or dissociation has been widely demonstrated^{3,4}. Such experiments are usually performed on weakly bound atoms that need to be stabilized by operating at cryogenic temperatures. Analogous experiments at room temperature⁵ are more difficult, because they require relatively strongly bound species that are not perturbed by random thermal fluctuations. But manipulation can still be achieved through electronic excitation of the atom or molecule by the electron current^{6–11} tunnelling between STM tip and surface at relatively high bias voltages^{10,11}, typically 1–5 V. Here we use this approach to selectively dissociate chlorine atoms from individual oriented chlorobenzene molecules adsorbed on a Si(111)- 7×7 surface. We map out the final destination of the chlorine daughter atoms, finding that their radial and angular distributions depend on the tunnelling current and hence excitation rate. In our system, one tunnelling electron has nominally sufficient energy to induce dissociation, yet the process requires two electrons. We explain these observations by a two-electron mechanism that couples vibrational excitation and dissociative electron attachment steps.

Figure 1 presents a sequence of STM images, acquired at room temperature and in ultrahigh vacuum, illustrating the dissociation of a single chlorobenzene molecule and the characterization of this event by careful STM imaging at different bias voltages. Figure 1a–d shows STM images obtained as the voltage applied to the Si(111)- 7×7 surface is increased from +1 V to +4 V. The chlorobenzene molecule undergoing dissociation at +4 V is labelled α ; it generates a chlorine atom bound to site β . γ and δ label chlorobenzene molecules that desorb at the same time. The experimental identification of the dissociation and desorption channels and of the dissociation fragments is based on the appearance of the relevant species in STM images as a function of the bias voltage¹², as illustrated in Fig. 1f–h. The desorption of molecules from sites γ and δ is apparent from the reappearance of the underlying silicon adatoms in Fig. 1g and h. The chlorine atom produced by single molecule dissociation appears as the bright feature at site β in

Fig. 1d–g and as the ‘missing’ adatom in Fig. 1h. Figure 1i and j provides a close-up of a molecular dissociation event, clearly showing the point at which the bright feature assigned to the chlorine atom appears under the tip while scanning from left to right across the surface at a bias voltage of +4 V and a tunnelling current of 100 pA.

The cross-sections of the chlorine atoms liberated from their chlorobenzene parent molecules are obtained by a method similar to that shown in Fig. 1, except that larger area scans are used, typically $350 \times 350 \text{ \AA}$. Using a dilute coverage of molecules ensures easy identification of the daughter–parent relationships, with tunnelling parameters chosen such that STM-induced diffusion of chlorine atoms¹³ is minimal (<1%).

The experimentally determined radial distribution of chlorine atoms is plotted in Fig. 2, clearly illustrating that the distributions differ when using tunnelling currents of 500 pA (Fig. 2a) or 100 pA (Fig. 2b). The data also show large separations between parent and daughter—up to 45 Å for a current of 500 pA (Fig. 2a), and up to 20 Å at 100 pA (Fig. 2b). Both values lie well beyond the nearest neighbour distance of 7.7 Å, so the chlorine atoms (or ions) are ejected from their parents with sufficient kinetic energy to travel a distance of several unit cells across the surface (an unambiguous example of hot atom transport as a result of bond dissociation¹⁴) until they are trapped and bound by a silicon adatom.

The configuration of the chemisorbed chlorobenzene molecule on the silicon surface is known^{15–17}; so provided that the azimuthal orientation of the chlorobenzene molecule can be determined, the influence of the azimuthal orientation of the carbon–chlorine bond on the angular distribution of ejected chlorine atoms can be established. As illustrated in Fig. 3a–f, high resolution STM imaging as a function of bias voltage can reveal which silicon adatom and rest-atom (that is, lower layer adatom) pair binds the molecule in its ‘di-σ’ bonding configuration on the surface. The chlorine atom is bound to one of the four carbon atoms that are not chemisorbed to

the silicon surface. Although we cannot identify the carbon atom carrying the chlorine atom, we are still able to map the angular distribution of chlorine atoms with respect to the bonding axes of the parent molecules within the limits of fourfold symmetry (see Fig. 3g inset).

Figure 3g and h plots the angular distribution of daughter chlorine atoms with respect to the adatom–rest-atom axis of the parent chlorobenzene molecule for tunnelling currents of 500 pA and 100 pA, respectively. The original direction of the carbon–chlorine bond lies at 60°. The data show that when using the higher tunnelling current, the chlorine atoms populate the complete range of angles: the orientation of the C–Cl bond is largely ‘forgotten’, even though the parent molecule is chemisorbed to the surface. We account for this broad angular distribution by assuming that chlorine atoms of higher energy are ejected at 500 pA and scattered by the silicon atoms that they encounter at radial distances up to ~45 Å (see Fig. 2a), before eventual capture by one of the silicon adatoms. At 100 pA (Fig. 3h), the angular distribution is anisotropic and peaks at 45–50°, which deviates slightly from the C–Cl bond axis at 60°. We attribute this to the observation that the daughter chlorine atoms are almost always found on silicon adatom sites^{12,18} (a very small number, <2%, are found at rest-atom sites); the adatom sites form a distinct crystallographic pattern around the parent chlorobenzene molecule, with an angle of 45–50° corresponding to the location of the silicon adatoms nearest to the parent molecule. It appears that capture of the ejected chlorine atoms by the first ring of silicon adatoms dominates the angular distribution at the lower tunnelling current, suggesting that the daughter atoms with relatively low kinetic energy are steered towards these nearest neighbour sites.

Figure 4a plots the measured rates of molecular dissociation and molecular desorption as a function of the tunnelling current. As shown previously¹⁶, the desorption rate exhibits a linear dependence on the current (the exponent is 0.9 ± 0.1), indicating that desorption is driven by a single tunnelling electron, specifically, by

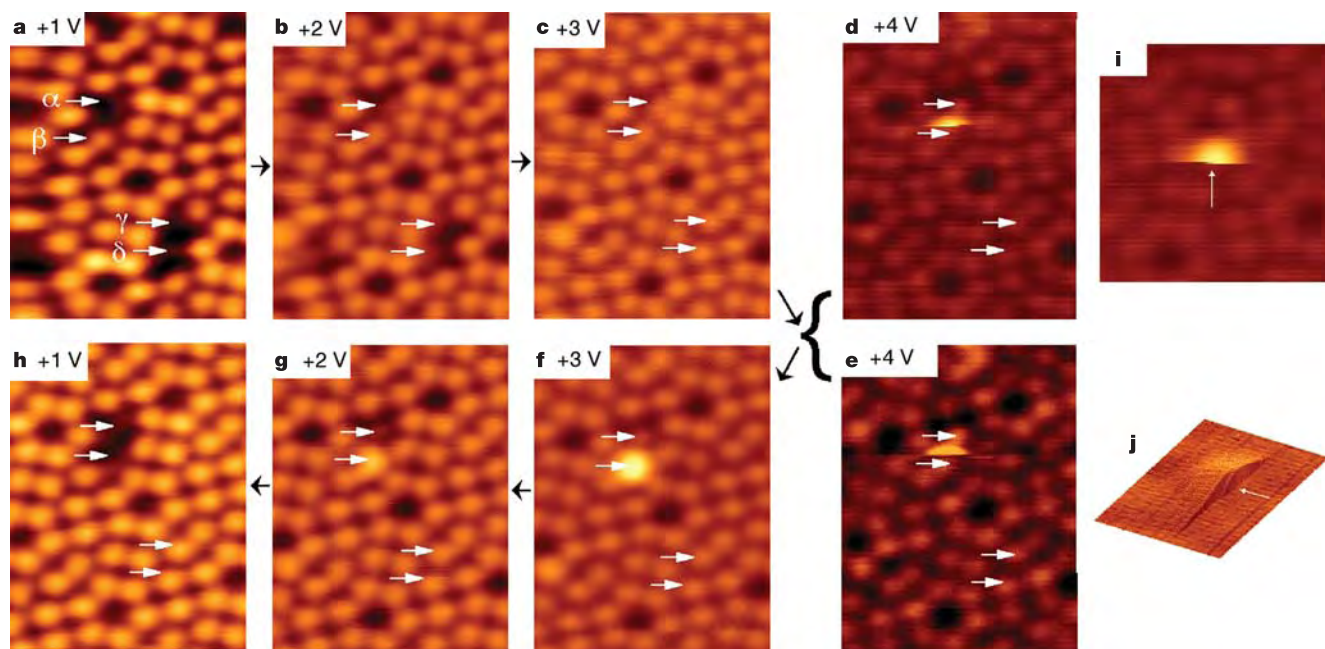


Figure 1 Dissociation of individual chlorobenzene molecules. STM images taken sequentially (see linking arrows) from chlorobenzene molecules chemisorbed on the Si(111)- 7×7 surface at room temperature as a function of sample bias voltage: **a**, +1 V, **b**, +2 V, **c**, +3 V, **d**, +4 V (rastered right to left), **e**, +4 V (left to right, obtained at the same time as **d**), **f**, +3 V, **g**, +2 V and **h**, +1 V; in each case the tunnelling current was

100 pA. The chlorobenzene molecule that undergoes dissociation at +4 V is labelled α ; it generates a chlorine atom bound to site β . γ and δ label chlorobenzene molecules that are desorbed. Panels **i** and **j** are, respectively, two-dimensional and three-dimensional STM images (+4 V, 100 pA, rastered left to right), showing the hallmark of molecular dissociation—generation of a chlorine atom that binds to a silicon adatom site.

the transient attachment of an electron to the antibonding π^* -state of the chemisorbed molecule (a negative-ion resonance¹⁹), causing vibrational excitation of the molecule–surface bond¹⁶. For negative bias voltages, hole attachment to the bonding π -states of the molecule mediates the desorption process similarly^{16,20}. In contrast, dissociation is a two-electron process: as the data in Fig. 4a show, the exponent of the current dependence of the dissociation rate is 1.8 ± 0.3 at +3.0 V.

Previous examples of STM-induced dissociation by vibrational excitation have found the number of electrons needed to achieve dissociation to equal the number expected on purely energetic considerations^{6,7,9,11,21}. In the case of chemisorbed chlorobenzene, *ab initio* calculations give an upper limit for the C–Cl chlorobenzene/Si(111)-7 $\times$ 7 bond energy of 1.9 eV (ref. 22), yet we find that two 3-eV electrons are needed to break the C–Cl bond even though each electron alone has enough energy to do so.

The dependence of the dissociation probability on the bias voltage (P.A.S. and R.E.P., manuscript in preparation) follows that for desorption¹⁶, and suggests that electron population of the antibonding π^* orbitals of the phenyl ring initiates both processes. Thus molecular desorption and carbon–chlorine bond dissociation are competing channels for the decay of the same negative-ion resonance state^{19,23} of the adsorbed molecule. However, whereas desorption occurs at both positive and negative sample bias voltages, we find no evidence for dissociation at negative bias, either for voltages between -0.5 V and -3 V at ~ 50 pA, or for tunnelling currents between 10 pA and 1 nA at -2 V. If dissociation resulted simply from (C–Cl) vibrational excitation, we would expect it to occur at both polarities, as does desorption; the fact

that this is not the case rules out a simple ‘vibrational heating’ mechanism. Chemical intuition suggests that dissociation might therefore involve the creation of a negative chlorine ion (that is, dissociative electron attachment, DEA^{24,25}); but the thermodynamic threshold for the creation of Cl^- is ~ 0 eV (ref. 26), raising the

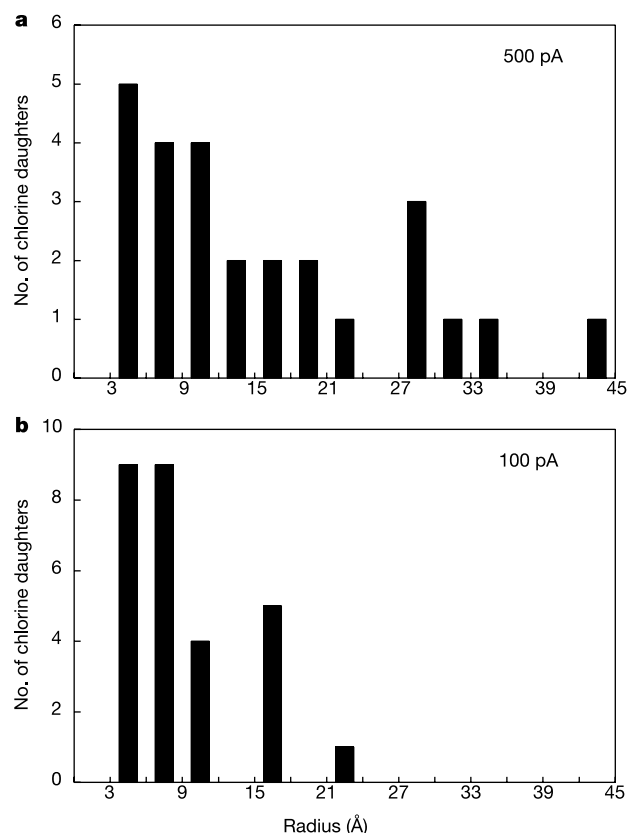


Figure 2 Radial distributions of daughter chlorine atoms. The number of daughter chlorine atoms as a function of radial distance (3 Å wide bins) from their parent chlorobenzene molecules on Si(111)-7 $\times$ 7, generated by tunnelling currents of 500 pA (a) and 100 pA (b).

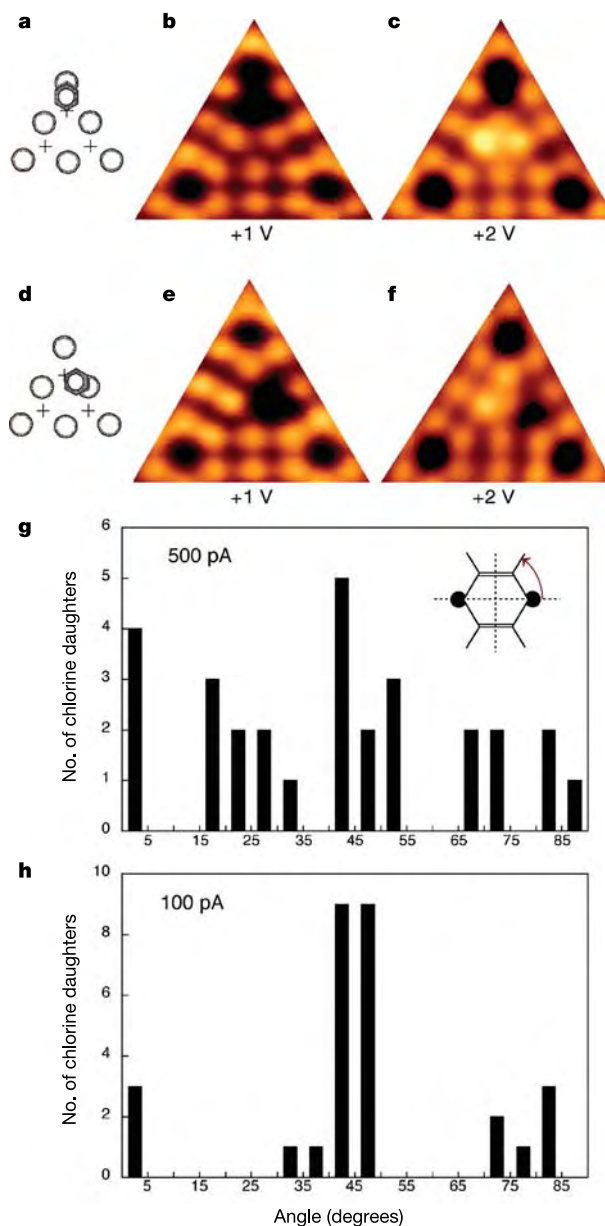


Figure 3 Angle-resolved dissociation. **a–f**, STM images and schematic diagrams (circles represent adatoms, crosses represent rest-atoms, and hexagons represent chlorobenzene molecules), showing the imaging characteristics of a single adsorbed chlorobenzene molecule as a function of sample bias (50 pA, bias voltages as marked). **a–c**, Chlorobenzene bonded to a corner adatom; **d–f**, chlorobenzene bonded to a centre adatom. Note the bright feature that appears over the bonding rest-atom in the case of a corner-bonded chlorobenzene molecule. This signature allows us to identify which of the two candidate rest-atoms is the bonding rest-atom when a chlorobenzene molecule is bonded to a centre adatom. **g**, Inset, diagram of chlorobenzene adsorbate (filled circles represent the bonding silicon atoms), showing planes of symmetry (dashed lines) and angular coordinate system (red arrow). **g**, **h**, Angular distribution (5° bins) of daughter chlorine atoms, relative to the adatom/rest-atom axis of each corresponding parent chlorobenzene molecule (as inset), generated by tunnelling currents of 500 pA (**g**) and 100 pA (**h**).

question of why we find that two electrons are needed to induce dissociation.

The cross-section for a DEA process depends very sensitively upon the symmetry of the intermediate negative-ion state. In gas phase experiments with chlorobenzene and other unsaturated halocarbon molecules, an electron is first captured into the π^* orbital of the molecule and then transfers into the σ^* orbital of the C–Cl bond to generate Cl^- ions^{26,27}. However, coupling of these two orbitals is forbidden under sp^2 symmetry (the C–Cl bond axis lies in the nodal plane of the π^* orbital), and only occurs via vibrational excitation of the out-of-plane C–Cl ‘wag’ mode²⁸ (vibrational frequency 61 meV; ref. 15)—that is, by vibronic coupling. In the gas phase, the same electron captured by the π^* orbital excites this

vibrational mode (losing energy in the process), and then transfers to the C–Cl σ^* orbital to generate Cl^- ions with near-zero kinetic energy²⁶. If the C–Cl bond does not reside in the π^* nodal plane, then there is no symmetry barrier and Cl^- ions are generated with considerable kinetic energy²⁶.

When chlorobenzene is chemisorbed on the Si(111)- 7×7 surface, the symmetry that the molecule had in the gas phase is perturbed. The aromatic nature of the ring is lost, and a cyclohexadiene-like, 2,5 di- σ bonded butterfly structure is formed. Two sp^3 carbon atoms bond to a pair of silicon atoms, but (crucially) a pair of sp^2 carbons, one of which bonds to the chlorine, is preserved on each wing^{15–17}; the wing of the molecule is a pseudo-chloroethene molecule, with sp^2 bonding. Gas-phase DEA of chloroethene is subject to the same $\pi^*-\sigma^*$ coupling constraints and proceeds via vibrational excitation, exactly as in chlorobenzene itself²⁶. We believe that the same vibronic mechanism for C–Cl dissociation must operate on the surface.

Adopting this mechanism to explain our observations still leaves the question of why two electrons are needed to induce dissociation in the adsorbed phase, whereas one electron suffices in the gas phase. The likely answer lies in the reduced lifetime of the negative-ion resonance in the chemisorbed chlorobenzene molecule. The gas-phase ‘symmetry forbidden’ electron transfer process is relatively slow, and takes ~ 100 fs (falling to ~ 20 fs without the symmetry barrier²⁹). The calculated lifetime of a positive-ion state of benzene in the butterfly geometry on the Si(100) surface (analogous to the chlorobenzene/Si(111)- 7×7 binding geometry) is ~ 7 – 20 fs (ref. 20). This lifetime calculation used, as a parameter, the STM-induced benzene desorption yield, a yield which is similar for both hole- and the electron-mediated chlorobenzene desorption¹⁶. It seems reasonable then that the lifetime of the negative-ion state of chemisorbed chlorobenzene is also ~ 7 – 20 fs, which is too short to allow the one-electron dissociation process to occur.

We propose that instead the first electron gives rise to vibrational excitation of the molecule, by way of temporary negative-ion formation (Fig. 4b). Some of the vibrationally highly excited molecules will desorb, but others will remain on the surface (Fig. 4c). The dependence of the radial and angular distributions of the ejected chlorine daughter atoms upon the tunnelling current (Fig. 2 and Fig. 3g and h) demonstrate that it is the time lapse (in the range of 100 ps to 1 ns) between the arrival of the first and second electrons that regulates the kinetic energy of the daughter. Some of this vibrational energy will reside in the C–Cl wag, as required for $\pi^*-\sigma^*$ coupling. If this mode is still excited when the second electron arrives (Fig. 4d), the π^* to σ^* barrier is relaxed and this electron can couple efficiently into the C–Cl group, resulting in dissociative electron attachment. The resulting daughter chlorine atom is captured by the surface (Fig. 4e). Note that if the C–Cl bond were bent out of the sp^2 plane, then $\pi^*-\sigma^*$ coupling would be allowed and a one-electron process could occur. This appears to be the case with STM dissociation of benzene on Cu(110) (ref. 21).

A vibrational lifetime for the C–Cl wag mode approaching 1 ns is not unreasonable: the lifetime of the Si–H stretch mode on Si(111), with frequency 258 meV, is 800 ps (ref. 30). The frequency of the C–Cl wag mode at 61 meV (ref. 15) will be decoupled from the surface phonons (≤ 62 meV; ref. 30) by the stretch and deformation modes of the (distorted) phenyl ring (~ 180 meV; ref. 15) and by the Si–C stretch (83 meV; ref. 17). In this scheme, the dependence of the chlorine atom’s kinetic energy on the time lapse between the two electron capture events arises because the π^* to σ^* transfer energy is more efficient if the C–Cl mode remains more highly excited. In the 500 pA experiment, the ‘second’ electron will encounter a more vibrationally excited molecule than in the 100 pA case (because the time lapse after the first electron scattering event is shorter by a factor of 5). The more efficient channelling of energy into the kinetic

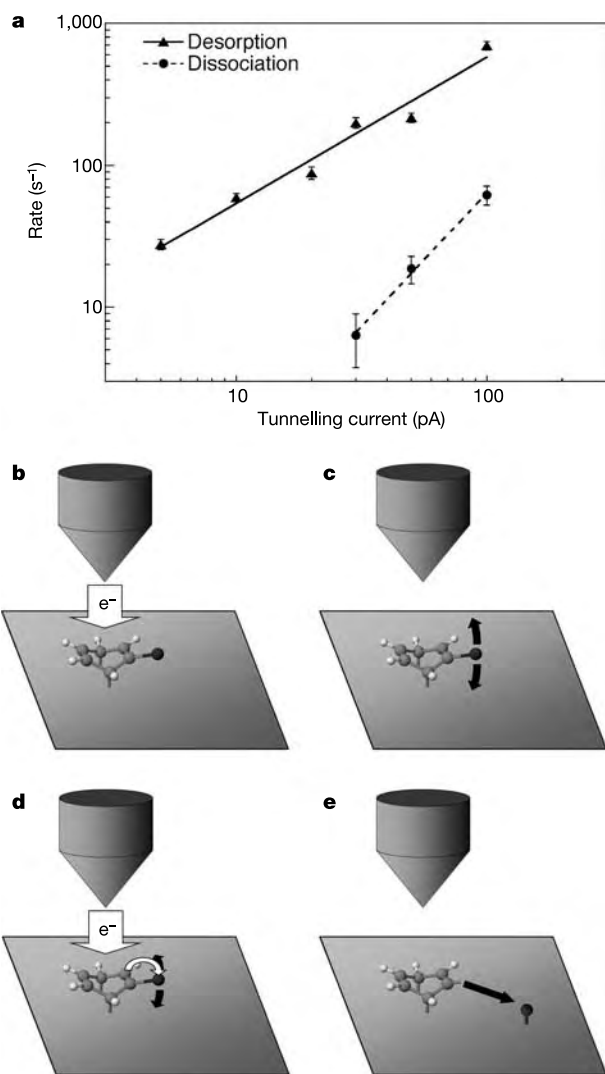


Figure 4 Dynamics of molecular dissociation. **a**, Rates of desorption (triangles) and of C–Cl bond dissociation (circles) for chlorobenzene on Si(111)- 7×7 as a function of the tunnelling current (I), with power-law fits, rate $\propto I^N$; for desorption, $N = 0.9 \pm 0.1$; for dissociation, $N = 1.8 \pm 0.3$ (sample bias voltage is +3 V in each case, error bars are 1 s.d. from the mean). **b–e**, Diagram of proposed vibrationally coupled two-electron mechanism (black atom is chlorine, black arrow is chlorine movement, and white arrow shows tunnelling electron). **b**, The first electron interacts with the chlorobenzene molecule; **c**, the molecule is left vibrationally excited (specifically, the C–Cl wag mode is excited); **d**, the second electron interacts with the molecule before the C–Cl wag mode has fully relaxed, leading to dissociation of the C–Cl bond by DEA; and **e**, capture of the ejected daughter Cl atom by the surface.

energy of the chlorine ion at 500 pA, compared with the 100 pA case, results in the more extended radial distribution and, because of surface scattering, the more isotropic angular distribution of the chlorine daughters. □

Received 7 December 2004; accepted 20 January 2005; doi:10.1038/nature03385.

- Eigler, D. M. & Schweizer, E. K. Positioning single atoms with a scanning tunnelling microscope. *Nature* **344**, 524–526 (1990).
- Bartels, L., Meyer, G. & Rieder, K.-H. Basic steps of lateral manipulation of single atoms and diatomic clusters with a scanning tunneling microscope tip. *Phys. Rev. Lett.* **79**, 697–700 (1997).
- Komeda, T., Kim, Y., Kawai, M., Persson, B. N. J. & Ueba, H. Lateral hopping of molecules induced by excitation of internal vibration mode. *Science* **295**, 2055–2058 (2002).
- Pascual, J. I., Lorente, N., Song, Z., Conrad, H. & Rust, H.-P. Selectivity in vibrationally mediated single-molecule chemistry. *Nature* **423**, 525–528 (2003).
- Fishlock, T. W., Oral, A., Egge, R. G. & Pethica, J. B. Manipulation of atoms across a surface at room temperature. *Nature* **404**, 743–745 (2000).
- Salam, G. P., Persson, M. & Palmer, R. E. Possibility of coherent multiple excitation in atom transfer with a scanning tunneling microscope. *Phys. Rev. B* **49**, 10655–10662 (1994).
- Stipe, B. C. *et al.* Single-molecule dissociation by tunneling electrons. *Phys. Rev. Lett.* **78**, 4410–4413 (1997).
- Lauhon, L. J. & Ho, W. Control and characterization of a multistep unimolecular reaction. *Phys. Rev. Lett.* **84**, 1527–1530 (2000).
- Hla, S.-W., Bartels, L., Meyer, G. & Rieder, K.-H. Inducing all steps of a chemical reaction with the scanning tunneling microscope tip: Towards single molecule engineering. *Phys. Rev. Lett.* **85**, 2777–2780 (2000).
- Lu, P. H., Polanyi, J. C. & Rogers, D. Electron-induced “localized atomic reaction” (LAR): chlorobenzene adsorbed on Si(111) 7 × 7. *J. Chem. Phys.* **111**, 9905–9907 (1999).
- Soukassian, L., Mayne, A. J., Carbone, M. & Dujardin, G. Atomic-scale desorption of H atoms from the Si(100)-2 × 1:H surface: inelastic electron interactions. *Phys. Rev. B* **68**, 035303 (2003).
- Boland, J. J. & Villarrubia, J. S. Formation of Si(111)-(1 × 1)Cl. *Phys. Rev. B* **41**, 9865–9870 (1990).
- Nakamura, Y., Mera, Y. & Maeda, K. Diffusion of chlorine atoms on Si(111)-(7 × 7) surface enhanced by electron injection from scanning tunneling microscope tips. *Surf. Sci.* **487**, 127–134 (2001).
- Brune, H., Wintterlin, J., Behm, R. J. & Ertl, G. Surface migration of “hot” adatoms in the course of dissociative chemisorption of oxygen on Al(111). *Phys. Rev. Lett.* **68**, 624–626 (1992).
- Cao, Y., Deng, J. F. & Xu, Q. Q. Stereo-selective binding of chlorobenzene on Si(111)-7 × 7. *J. Chem. Phys.* **112**, 4759–4767 (2000).
- Sloan, P. A., Hedouin, M. F. G., Palmer, R. E. & Persson, M. Mechanism of molecular manipulation with the scanning tunneling microscope at room temperature: chlorobenzene/Si(111)-(7 × 7). *Phys. Rev. Lett.* **91**, 118301 (2003).
- Li, Z.-H., Li, Y.-C., Wang, W.-N., Cao, Y. & Fan, K.-N. A density functional theory study on the adsorption of chlorobenzene on the Si(111)-7 × 7 surface. *J. Phys. Chem. B* **108**, 14049–14055 (2004).
- Jensen, J. A., Yan, C. & Kummel, A. C. Direct chemisorption site selectivity for molecular halogens on the Si(111)-(7 × 7) surface. *Phys. Rev. Lett.* **76**, 1388–1391 (1996).
- Palmer, R. E. & Rous, P. J. Resonances in electron scattering by molecules on surfaces. *Rev. Mod. Phys.* **64**, 383–440 (1992).
- Alavi, S. *et al.* Inducing desorption of organic molecules with a scanning tunneling microscope: theory and experiments. *Phys. Rev. Lett.* **85**, 5372–5375 (2000).
- Komeda, T., Kim, Y., Fujita, Y., Sainoo, Y. & Kawai, M. Local chemical reaction of benzene on Cu(110) via STM-induced excitation. *J. Chem. Phys.* **120**, 5347–5352 (2004).
- Jiang, G., Polanyi, J. C. & Rogers, D. Electron and photon irradiation of benzene and chlorobenzene on Si(111) 7 × 7. *Surf. Sci.* **544**, 147–161 (2003).
- Martin, F. *et al.* DNA strand breaks induced by 0–4 eV electrons: the role of shape resonances. *Phys. Rev. Lett.* **93**, 068101 (2004).
- Dixon-Warren, St.-J., Jensen, E. T. & Polanyi, J. C. Direct evidence for charge-transfer photodissociation at a metal surface: CCl₄/Ag(111). *Phys. Rev. Lett.* **67**, 2395–2398 (1991).
- Palmer, R. E. Electron-molecule dynamics at surfaces. *Prog. Surf. Sci.* **41**, 51–108 (1992).
- Dressler, R., Allan, M. & Haselbach, E. Symmetry control in bond cleavage processes: dissociative electron attachment to unsaturated halocarbons. *Chimia* **39**, 385–389 (1985).
- Modelli, A. Electron attachment and intramolecular electron transfer in unsaturated chloroderivatives. *Phys. Chem. Chem. Phys.* **5**, 2923–2930 (2003).
- Skalicky, T., Chollet, C., Pasquier, N. & Allan, M. Properties of the π* and σ* states of the chlorobenzene anion determined by electron impact spectroscopy. *Phys. Chem. Chem. Phys.* **4**, 3583–3590 (2002).
- Fontanesi, C., Baraldi, P. & Marcaccio, M. On the dissociation dynamics of the benzyl chloride radical anion. An ab initio dynamic reaction coordinate analysis study. *J. Mol. Struct. (Theochem)* **548**, 13–20 (2001).
- Guyot-Sionnest, P., Dumas, P., Chabal, Y. J. & Higashi, G. S. Lifetime of an adsorbate-substrate vibration: H on Si(111). *Phys. Rev. Lett.* **64**, 2156–2159 (1990).

Acknowledgements We thank the EPSRC and the European Research Training Networks ‘Manipulation of individual atoms and molecules’ and AMMIST for support. P.A.S. acknowledges studentship support from the School of Physics and Astronomy and from EPSRC. We also thank J. C. Polanyi for pointing us in the direction of vibrationally activated chemical reactions.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.E.P. (r.e.palmer@bham.ac.uk).

Subducted banded iron formations as a source of ultralow-velocity zones at the core–mantle boundary

David P. Dobson & John P. Brodholt

Department of Earth Sciences, University College London, Gower Street, London, WC1E 6BT, UK

Ultralow-velocity zones (ULVZs) are regions of the Earth’s core–mantle boundary about 1–10 kilometres thick exhibiting seismic velocities that are lower than radial-Earth reference models by about 10–20 per cent for compressional waves and 10–30 per cent for shear waves. It is also thought that such regions have an increased density of about 0–20 per cent (ref. 1). A number of origins for ULVZs have been proposed, such as ponding of dense silicate melt², core–mantle reaction zones³ or underside sedimentation from the core⁴. Here we suggest that ULVZs might instead be relics of banded iron formations subducted to the core–mantle boundary between 2.8 and 1.8 billion years ago. Consisting mainly of interbedded iron oxides and silica, such banded iron formations were deposited in the world’s oceans during the late Archaean and early Proterozoic eras. We argue that these layers, as part of the ocean floor, would be recycled into the Earth’s interior by subduction⁵, sink to the bottom of the mantle and may explain all of the observed features of ULVZs.

Between 2.8 and 1.8 Gyr ago, the Earth’s oceans underwent several periods of oxidation, which are commonly thought to be related to the appearance of photosynthesis in bacteria⁶. Before these oxidation events, ocean waters contained large quantities of (soluble) ferrous iron as a result of chemical weathering of the proto-continents and leaching of the ocean floor near volcanic centres. As primitive life in the form of blue-green algae took hold, oxygen released by photosynthesis would have reacted with the ferrous iron to form magnetite- and haematite-rich sedimentary precipitates: the banded iron formations (BIFs) (Fig. 1). BIFs are the world’s most important source of iron, with an estimated 4×10^{14} t globally of 2.5–2.2-billion-year-old BIFs⁷. The most commonly preserved BIFs are interpreted as having been deposited in relatively shallow marine facies, resulting from the interaction between oxidized surface waters and upwelling Fe-rich, reduced deep ocean water, and can be several hundred metres thick. It is likely, however, that a permanent redox front would have existed between oxidized surface waters and reduced iron-rich deep waters in the ocean basins as a whole, not just on continental shelves. Magnetite would precipitate at this front and rain onto the ocean floor forming deep-water BIFs. As part of the ocean floor, these BIFs would then have been recycled into the Earth’s interior by subduction⁵ (Fig. 2).

The high concentration of iron oxide in BIFs (up to 85 wt%) (ref. 8) results in a density that is greater than any other part of the slab or upper and lower mantle (see below). Subduction processes were certainly occurring at least by the late Archaean, as evidenced by 3.7–2.5-billion-year-old occurrences of boninites and ophiolites^{9,10}. Indeed, late Archaean greenstone belts that are interpreted as ophiolites commonly preserve BIFs in the accretionary prism and are interbedded with basaltic magmatism^{9,11}. Seismic tomography studies¹² demonstrate that subducted slab material can currently penetrate to the base of the mantle. A similar range of subduction styles are likely to have operated below these ancient subduction zones, transporting at least some BIFs to the core–mantle boundary (CMB).

For subducted BIFs to constitute the ULVZs, a number of physical and chemical criteria must be met. First, BIFs must be negatively buoyant in the mantle. Second, the BIF material must be

able to segregate onto the CMB. Third, BIFs must not react with or dissolve into the core or mantle over the past 3 Gyr. Fourth, the physical properties of BIFs under lower mantle conditions must match those of the ULVZs. And finally, the volume of deep-water BIFs produced and subducted must agree approximately with the volumes of the ULVZs.

The density of a BIF made up of 50% Fe_3O_4 (as magnetite) and 50% SiO_2 (in the quartz, stishovite and columbite phases as appropriate) will be about 3.9 g cm^{-3} on the ocean floor, 5.1 g cm^{-3} at the bottom of the upper mantle (at 2,000 K and 25 GPa) and 6.6 g cm^{-3} in the CMB (at 3,000 K and 136 GPa) (refs 13, 14). Because the oxidation state of the mantle is nearer to iron-wustite than magnetite, it is likely that the Fe_3O_4 component will be reduced to FeO. This would have the effect of increasing the density at the base of the upper mantle to 5.4 g cm^{-3} and to 8.1 g cm^{-3} in the CMB^{15,16}. Regardless of the exact phase (NaCl-, NiAs- or inverse (i)NiAs-structured) or oxidation state, the density of the subducting BIF is always much higher than the ambient mantle. At the CMB, the density of a 50% SiO_2 , 50% FeO BIF ($\sim 8\text{ g cm}^{-3}$) is higher than the lower mantle (5.6 g cm^{-3}) but less than the outer core (10 g cm^{-3}) (ref. 17), and BIFs are therefore neutrally buoyant between the outer core and lower mantle.

It seems likely, therefore, that BIFs could have been transported into the deep lower mantle by the subduction conveyor belt and that they would tend to sink towards the CMB owing to their high density. In order to be concentrated on the CMB, BIFs would still need to separate from the 100 km or so of silicate slab component. The exact details of this separation are likely to be highly complex. However, one can envisage two simplified end-member processes. In the first, BIF material falls through the underlying slab as diapirs of similar size to the BIF thickness. We can make a conservative estimate that the maximum thickness of BIF deposits preserved on the proto-continents represents the mean depth of BIF deposited in

the ocean deeps, about 500 m. A 500 m BIF body will take around 1 Gyr to fall through 100 km of slab with a viscosity of 10^{21} Pa s . This requires that the slab remains at the CMB for about 1 Gyr. The timescales for this would be significantly reduced, however, if the BIF was thickened—owing to, for example, the increased resistance as the slab entered the lower mantle, or mass conserving flow as the slab deforms in response to subduction from a spherical surface. Some models explaining the anomalous Re-Os isotope systematics of ocean island basalts invoke slab material residing at the CMB for billions of years¹⁸, so this timescale for BIF settling may be reasonable. Alternatively, if the slab arrives at the CMB at a steeper angle than about 10° then weak BIF material could separate by flowing along the surface of the slab. If we assume laminar flow of BIF between the slab and overlying mantle, a 500-m-thick layer could settle at 1 cm yr^{-1} if it had a viscosity of 10^{18} Pa s (Supplementary Fig. 1). Once again, increasing the layer thickness enhances the settling rate. This could either cause BIF layers to locally thicken and enhance diapiric settling, or be a direct method for separating the BIFs onto the CMB.

The question then is what would happen to BIFs during subduction and whether they could remain at the CMB for 2 to 3 Gyr. First of all, laser-heated diamond cell measurements¹⁹ show that the FeO melting temperature at 100 GPa is above 5,000 K and FeO will be solid under all mantle conditions. Any reaction with the mantle would, therefore, have to be via a solid-state reaction. The time for a subducting slab to reach the CMB, even if subducting very slowly at 1 cm yr^{-1} , is 300 Myr, sufficient only to diffuse Fe 100 m or so. Even over 3 Gyr, diffusion can only interchange Fe and Mg by about 0.5 km so any sort of solid-state exchange of Fe into the surrounding mantle will not dilute the iron content by any appreciable amount. Assuming BIF arrives at the CMB, will it react or dissolve into the outer core? At the moment the answer to this is not known. If, as has been recently proposed²⁰, the core formed in oxygen-fugacity equilibrium with the mantle, then the outer core would become more saturated with oxygen with time as the inner core grew, and an FeO component at the CMB could remain stable for the lifetime of the Earth. If, however, the core was undersaturated in oxygen, the FeO component of the BIF would be soluble in the liquid outer core. Even if this were the case, it is not clear that all the FeO would be extracted as dissolution of FeO into the core would leave a residue of (probably) insoluble SiO_2 . The FeO-depleted residue could become concentrated at the base of the BIF layer and effectively chemically



Figure 1 A 2.5-Gyr-old banded iron formation from Hamersley, Australia. The grey bands predominantly contain the iron oxide haematite and the brown bands are rich in siliceous chert.

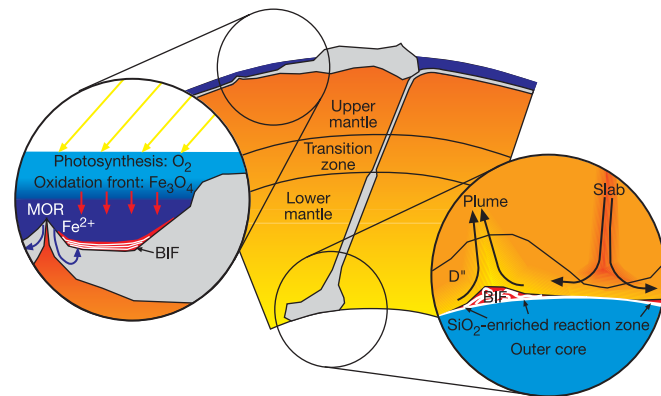


Figure 2 Schematic cross-section of the Earth's mantle showing formation of BIFs in the oceans and segregation at the core–mantle boundary. Left inset, early life in the photic zone of the oceans releases oxygen from the atmosphere by photosynthesis. This reacts with Fe^{2+} in solution in deep ocean waters to produce insoluble iron oxides, which are precipitated as BIF on the ocean floors. BIF is recycled into the mantle at subduction zones. MOR, mid-ocean ridge. Right inset, BIF ponds at the core–mantle boundary, where it is neutrally buoyant. Convection currents concentrate BIF below regions of mantle upwelling where the ultralow-velocity zone is now observed.

isolate the overlying, unreacted BIF. Any reaction zone between the core and mantle is limited to the capillary rise height of the outer core liquid, which is 20 m (ref. 21). Thus a 20-m-thick layer of SiO₂ at the base of any ponded BIF would be sufficient to stop further reaction with the outer core. However, as mentioned above, if the core is saturated or in equilibrium with oxygen, there is no need for such an explanation.

We can now consider whether the physical properties of BIFs are consistent with observations of the ULVZs. The high-pressure/high-temperature seismic properties of the high-pressure FeO phases are not known at present, so we cannot compare directly with the observed reductions in the compressional wave velocity (v_p) and shear wave velocity (v_s). We can, however, estimate the bulk sound velocity (v_ϕ) from the bulk modulus (K) and density (ρ) using the equation ($v_\phi = (K/\rho)^{1/2}$) and compare that to mantle values derived from spherical averaged velocity profiles such as PREM¹⁷. Pure FeO takes on the NiAs (or iNiAs) structure under CMB conditions¹⁶, but recent experiments²² suggest that small amounts of MgO can stabilize the low-pressure, NaCl structure. On the basis of the high-pressure equations of state of NiAs- and NaCl-FeO, and stishovite^{14–17}, we estimate that the bulk sound velocity of BIF will be $9.3 \pm 0.4 \text{ km s}^{-1}$, which is 10% slower than PREM values for the base of the mantle. Although the bulk sound velocity is only a rough indicator, not taking into account the shear moduli of the relevant phases, this value is consistent with the seismically observed properties of the ULVZs. Indeed, if BIFs are in the NaCl structure, the recently observed shear-mode softening in FeO (ref. 23) associated with the NaCl-rhombohedral transition will cause a large decrease in v_s relative to v_p , which is one of the most striking properties of the ULVZs.

The final question is whether enough BIFs were produced to account for the observed volume of ULVZs. Again this can only be a rough comparison because first, ULVZs do not appear to be global features in the D'' layer nor has all of D'' been sampled yet, and second, there are no preserved deep-water BIFs (presumably they have been subducted). However, our conservative estimate of deep-ocean BIF thickness, 500 m, would produce a 1.3-km-thick layer if spread evenly along the core–mantle boundary. This is consistent with ULVZ thicknesses of about 5–20 km, over about 12% of the surface of the CMB (ref. 24).

There are still a number of unanswered questions—in particular, can we concentrate thin layers of BIF (~500 m) into layers of a few kilometres thick, consistent with ULVZs, or would it be stretched into even thinner seismically invisible layers? Viscous drag from the base of the mantle would deform weak BIF (or its post-BIF phases) causing it to become concentrated at regions of mantle upwelling, because its gravitational stability at the CMB would inhibit significant entrainment into mantle plumes. In other words, it may pond beneath regions of upwelling, which is consistent with the observation of ULVZs below seismically slow regions of the deep lower mantle¹.

We can now ask the question: what would be the consequences for the dynamics of the CMB if subducted BIF does constitute the ULVZs? FeO is expected to be either a very small-bandgap semiconductor or metallic at mantle pressures and temperatures^{18,25}, with a large electrical conductivity around 10^4 – 10^6 S m^{-1} . A 1 km average thickness of subducted BIF would therefore contribute between 10^7 and 10^9 S of conductance to the base of the lower mantle. A laterally varying, highly conductive layer at the base of the mantle, such as that provided by subducted BIF, has been invoked to explain phenomena such as nutations²⁶ and core–mantle coupling²⁷. Small changes in the Earth's length of day that occur on a decadal timescale are attributed to momentum transfer between the core and mantle—core–mantle coupling. One way of achieving this coupling is through interactions of a highly conductive layer at the base of the lower mantle (with 10^8 S of conductance) with the electromagnetic field generated in the core²⁸. Normal mantle is

insufficiently conductive to be a source of electromagnetic coupling, and silicate or metallic melts have been ruled out²¹ as the conductive layer. An ULVZ consisting of the deep BIF assemblage would, however, have the correct properties to explain core–mantle coupling and the other phenomena mentioned above.

The origin of ULVZs remains unknown. We have put forward a hypothesis that may explain all the observed features, although we need further experimental and theoretical data in order to test it. The absence of extensively preserved ocean floor BIFs is explicable in terms of their subduction along with the underlying oceanic crust. We believe that the most likely ultimate fate of subducted BIF is ponding at the core–mantle boundary and that the properties of a deep BIF assemblage can explain observations of ULVZs. One of the most enigmatic features of the deepest interior of the Earth may therefore owe its existence to some of the earliest and most primitive life. □

Received 16 July 2004; accepted 28 January 2005; doi:10.1038/nature03430.

- Garnero, E. J., Revenaugh, J., Williams, Q., Lay, T. & Kellogg, L. H. in *The Core–Mantle Boundary Region* (eds Gurnis, M., Wyession, M. E., Knittle, E. & Buffet, B. A.) 319–334 (American Geophysical Union, Washington DC, 1998).
- Williams, Q. & Garnero, E. J. Seismic evidence for partial melt at the base of the mantle. *Science* **273**, 1528–1530 (1996).
- Knittle, E. & Jeanloz, R. Simulating the core–mantle boundary: an experimental study of high-pressure reactions between silicates and liquid iron. *Geophys. Res. Lett.* **16**, 609–612 (1989).
- Buffett, B. A., Garnero, E. J. & Jeanloz, R. Sediments at the top of Earth's core. *Science* **290**, 1338–1342 (2000).
- Polat, A., Hofman, A. W. & Rosing, M. T. Boninite-like volcanic rocks in the 3.7–3.8 Ga Isua greenstone belt, West Greenland: geochemical evidence for intra-oceanic subduction zone processes in the early Earth. *Chem. Geol.* **184**, 231–254 (2002).
- Cloud, P. Paleogeological significance of the banded iron formation. *Econ. Geol.* **68**, 1135–1143 (1973).
- James, H. L. in *Iron Formation: Facts and Problems* (eds Trendall, A. F. & Morris, R. C.) 472–473 (Elsevier, Amsterdam, 1983).
- Castro, L. O. Genesis of banded iron-formations. *Econ. Geol.* **89**, 1384–1397 (1994).
- Kusky, T. M., Li, J.-H. & Tucker, R. D. The Archean Dongdangzi Ophiolite Complex, North China Craton: 2.505-billion-year-old oceanic crust and mantle. *Science* **292**, 1142–1146 (2001).
- Gnaneshwar Rao, T. & Naqvi, S. M. Geochemistry, depositional environment and tectonic setting of the BIFs of the late Archean Chitradurga Schist Belt, India. *Chem. Geol.* **121**, 217–243 (1995).
- Taira, A., Pickering, K. T., Windley, B. F. & Soh, W. Accretion of Japanese island arcs and implications for the origin of Archean greenstone belts. *Tectonics* **11**, 1224–1244 (1992).
- Káráson, H. & Van der Hilst, R. D. in *The History and Dynamics of Global Plate Motion* (eds Richards, M. R., Gordon, R. & Van der Hilst, R. D.) 277–288 (American Geophysical Union, Washington DC, 2000).
- Mao, H. K., Takahashi, T., Basset, W. A., Kinsland, G. L. & Merrill, L. Isothermal compression of magnetite up to 320 kbar and pressure induced phase transformation. *J. Geophys. Res.* **79**, 1165–1170 (1974).
- Ross, N. L., Shu, J. F., Hazen, R. M. & Gasparik, T. High-pressure crystal chemistry of stishovite. *Am. Mineral.* **75**, 739–747 (1990).
- Will, G., Hinz, E. & Nuding, W. The compressibility of FeO measured by energy dispersive X-ray diffraction in a diamond anvil squeezer up to 200 kbar. *Phys. Chem. Miner.* **6**, 157–167 (1980).
- Fei, Y. W. & Mao, H. K. In-situ determination of the NiAs phase of FeO at high-pressure and temperature. *Science* **266**, 1678–1680 (1994).
- Dziewonski, A. M. & Anderson, D. L. Preliminary reference Earth model. *Phys. Earth Planet. Inter.* **25**, 297–356 (1981).
- Hauri, E. H. & Hart, S. R. Re-Os isotope systematics of HIMU and EMII oceanic island basalts from the south-pacific ocean. *Earth Planet. Sci. Lett.* **114**, 353–371 (1993).
- Knittle, E. & Jeanloz, R. The high pressure phase diagram of Fe_{0.94}O: a possible constituent of the Earth's core. *J. Geophys. Res.* **96**, 16169–16180 (1991).
- Rubie, D. C., Gessmann, C. K. & Frost, D. J. Partitioning of oxygen during core formation on the Earth and Mars. *Nature* **429**, 58–61 (2004).
- Poirier, J. P., Malavergne, V. & Le Mouél, J. L. in *The Core–Mantle Boundary Region* (eds Gurnis, M., Wyession, M. E., Knittle, E. & Buffet, B. A.) 131–138 (American Geophysical Union, Washington DC, 1998).
- Kondo, T., Ohtani, E., Hirao, N., Yagi, T. & Kikegawa, T. Phase transitions of (Mg,Fe)O at megabar pressures. *Phys. Earth Planet. Inter.* **143–144**, 201–213 (2004).
- Jacobsen, S. D., Spetzler, H., Reichmann, H. J. & Smyth, J. R. Shear waves in the diamond-anvil cell reveal pressure-induced instability in (Mg,Fe)O. *Proc. Natl Acad. Sci. USA* **101**, 5867–5871 (2004).
- Persch, S. E. & Vidale, J. E. Reflection properties of the core–mantle boundary from global stacks of PcP and ScP. *J. Geophys. Res.* **109**, doi:10.1029/2003JB002768 (2004).
- Knittle, E. & Jeanloz, R. High-pressure metallization of FeO and implications for the Earth's core. *Geophys. Res. Lett.* **13**, 1541–1544 (1986).
- Buffett, B. A. Constraints on magnetic energy and mantle conductivity from the forced nutations of the Earth. *J. Geophys. Res.* **97**, 19581–19597 (1992).
- Stix, M. & Roberts, P. H. Time-dependent electromagnetic core–mantle coupling. *Phys. Earth Planet. Inter.* **36**, 49–60 (1984).
- Holme, R. Electromagnetic core–mantle coupling - I. Explaining decadal changes in the length of day. *Geophys. J. Int.* **132**, 167–180 (1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank P. Bown and M. Kendall for discussions. This work was performed under the auspices of the Deep Earth System Consortium. D.D. was supported by a Royal Society University Research Fellowship.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.P.D. (d.dobson@ucl.ac.uk).

'Lophenteropneust' hypothesis refuted by collection and photos of new deep-sea hemichordates

Nicholas D. Holland¹, David A. Clague², Dennis P. Gordon³,
Andrey Gebruk⁴, David L. Pawson⁵ & Michael Vecchione⁶

¹Marine Biology Research Division, Scripps Institution of Oceanography (UCSD), La Jolla, California 92093-0202, USA

²Monterey Bay Aquarium Research Institute, 7700 Sandholdt Road, Moss Landing, California 95039-9644, USA

³National Institute of Water and Atmospheric Research, PO Box 14-901, Kilbirnie, Wellington, New Zealand

⁴P. P. Shirshov Institute of Oceanology, Russian Academy of Sciences, Nakhimovsky Prospect 36, Moscow 117851, Russia

⁵National Museum of Natural History, Smithsonian Institution, Washington, DC 20013-7012, USA

⁶NOAA National Systematics Laboratory, National Museum of Natural History, Washington, DC 20013-7012, USA

The deep ocean is home to a group of broad-collared hemichordates—the so-called 'lophenteropneusts'—that have been photographed gliding on the sea floor^{1–8} but have not previously been collected. It has been claimed that these worms have collar tentacles and blend morphological features of the two main hemichordate body plans, namely the tentacle-less enteropneusts and the tentacle-bearing pterobranchs. Consequently, lophenteropneusts have been invoked as missing links to suggest that the former evolved into the latter⁵. The most significant aspect of the lophenteropneust hypothesis is its prediction that the fundamental body plan within a basal phylum of deuterostomes was enteropneust-like. The assumption of such an ancestral state influences ideas about the evolution of the vertebrates from the invertebrates^{9–14}. Here we report on the first collected specimen of a broad-collared, deep-sea enteropneust and describe it as a new family, genus and species. The collar, although disproportionately broad, lacks tentacles. In addition, we find no evidence of tentacles in the available deep-sea photographs (published and unpublished) of broad-collared enteropneusts, including those formerly designated as lophenteropneusts. Thus, the lophenteropneust hypothesis was based on misinterpretation of deep-sea photographs of low quality and should no longer be used to support the idea that the enteropneust body plan is basal within the phylum Hemichordata.

The recently collected enteropneust (Figs 1a, b and 2) is described below as a new family, genus and species in the class Enteropneusta of the phylum Hemichordata.

Diagnoses. *Torquaratoridae* fam. nov.: proboscis and collar each conspicuously broader from side-to-side than in their other dimensions (anteroposterior and dorsoventral); with prominent hepatic caeca, but lacking synapticles.

Torquarator gen. nov.: diagnosis as for family.

Torquarator bullocki n. sp. Description: Living adult 70 mm long and 15 mm wide through collar (smallest length-to-width ratio

known for any adult enteropneust). Northeastern Pacific. Colour in life tan anteriorly, grading into light blue posteriorly, except where large white oocytes and dark grey gut contents show through translucent body wall. Proboscis a low dome with breadth (8 mm) conspicuously exceeding anteroposterior or dorsoventral dimensions (both about 5 mm); includes buccal diverticulum (stomochord) and proboscis skeleton with very short anterior and posterior horns; proboscis base encircled by basiepidermal nerve ring. Collar breadth (15 mm) greater than anteroposterior or dorsoventral dimensions (both about 7 mm); wide mouth at anterior end of collar opens into spacious buccal cavity; collar with midventral slit (Fig. 1b, black arrow), paired periahaemal spaces, and collar nerve cord lacking lumen. Trunk accompanied along entire length by lateral wings (sheet-like folds of body wall) and by dorsal and ventral trunk nerves. Lateral wings curling over dorsal surface of trunk in life, but retracting after fixation. Along anterior third of trunk (Fig. 1b, gt), lateral wings include several hundred separate ovaries containing oocytes of diverse sizes; largest oocytes (about 0.5 mm in diameter) white and visible through body wall (Fig. 1a, b). Pharynx (running through anterior 60% of gonadal region) dorsoventrally flattened without subdivision into respiratory region dorsally and digestive region ventrally. Several dozen pharyngeal gill slits in anteroposterior row on either side of dorsal midline with corresponding gill pores in overlying epidermis. No synapticles joining primary and secondary gill bars. Oesophagus (posterior 40% of gonadal region) subdivided into anterior and posterior zones (thin-walled, dorsoventrally flattened) separated by middle zone (thick-walled, circular in cross-section). Intestine, traversing posterior two-thirds of trunk (Fig. 1b, it) as long hepatic region anteriorly (pleated on either side by several dozen hepatic caeca with overlying epidermis closely following their contours) followed by short posterior region (lacking caeca or obvious anal sphincter). In all body regions, musculature poorly developed; epidermal mucus cells uniformly abundant.

Etymology. *Torquarator* derives from the Latin *torques* (collar) plus *arator* (ploughman) and refers to sediment-harvesting by the collar region. The specific name, *bullocki*, honours Professor Theodore Holmes Bullock, who published his PhD dissertation on

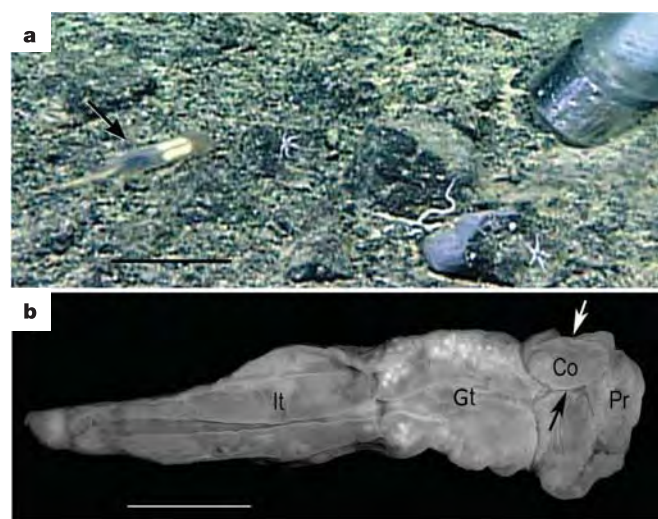


Figure 1 Holotype of *Torquarator bullocki* (Phylum Hemichordata, Class Enteropneusta). **a**, Living specimen (arrowed) crawling on deep-sea floor just before collection by a hose suction sampler (at right). Scale bar, 5 cm. **b**, Ventral view of the same specimen after collection and fixation showing the proboscis (Pr), collar (Co), gonadal trunk region (Gt) and intestinal trunk region (It); the black arrow indicates the midventral slit in the collar, and the white arrow the artefactual compression of the right side of the collar that occurred during capture and fixation. Scale bar, 1 cm.

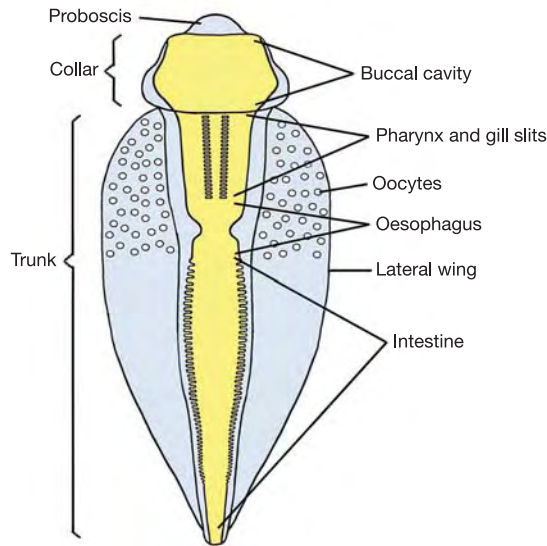


Figure 2 A diagram of *Torquarator bullocki* in dorsal view. The lateral wings are shown unfolded from the dorsal side and extended laterally to permit a clearer view of the course of the digestive tract (in yellow).

enteropneusts in 1940 and has been interested in them ever since. **Holotype.** The holotype (an adult female) was discovered on 27 July 2002, living at a depth of 1,901 m in the northeastern Pacific (42.5° N, 126.7° W) at an oxygen concentration of 1.45 ml l^{-1} , a salinity of 34.4‰, and a temperature of 2.01 °C. The living enteropneust (Fig. 1a) was videotaped for about 10 min, collected with a hose suction sampler, brought to the surface and preserved at once in 4% formalin in sea water. The fixation was appropriate for histology but precluded molecular phylogenetic study. We photographed the intact specimen and then prepared it as 20- μm serial paraffin sections stained with haematoxylin–eosin–toluidine blue.

The sectioned holotype is deposited at the Scripps Institution of Oceanography Benthic Invertebrate Collection as specimen number SIO-BICH2.

The exceptional breadth of the proboscis and collar distinguish the Torquaratoridae from all other enteropneust families (Ptychoderidae, Spengelidae, Harrimaniidae and Saxipendidiidae^{15,16}). Analysis of additional characters shows that the Torquaratoridae does not simply represent a previously recognized family with the collar adaptively broadened to facilitate the gathering of deep-sea sediments. Whereas the Torquaratoridae and the Ptychoderidae differ from all other enteropneusts in having prominent hepatic caeca, the former family differs from the latter in lacking synaptics. The family-level status of additional broad-collared deep-sea enteropneusts (Fig. 3, Table 1) remains undetermined until they can be collected.

When observed alive, the holotype of *T. bullocki* was gliding smoothly forwards at about 5 mm min^{-1} , the movement presumably mediated by the coordinated beating of epidermal cilia. While crawling, the worm did not leave behind an obvious fecal trail on the ocean floor, although other deep-living enteropneusts often produce such trails in a meandering (Fig. 3a) and/or spiral pattern (Fig. 3b, e). The ventral (mouth) side was oriented towards a hard substratum of fragmented volcanic glass covered with a very thin layer of soft sediment. The poorly developed body muscles of *T. bullocki* indicate that this species cannot burrow, but it might periodically swim or float above the bottom, as has been observed for another wide-collared deep-sea enteropneust⁸.

The oocyte size and the gut contents permit some inferences about aspects of the reproductive and feeding biology, respectively. It is likely that *T. bullocki*, as is typical of other deep-sea invertebrates with ripe oocytes on the order of 0.5 mm in diameter, produces larvae that are lecithotrophic and develop directly¹⁷. The contents of the intestine include finely granular material, long filamentous bacteria, sponge spicules, holothurian ossicles, diatoms, coccolithophores and benthic foraminiferans. The foraminiferans contain remnants of cytoplasm and may, with the bacteria, constitute a significant energy source for the enteropneust. In addition, the

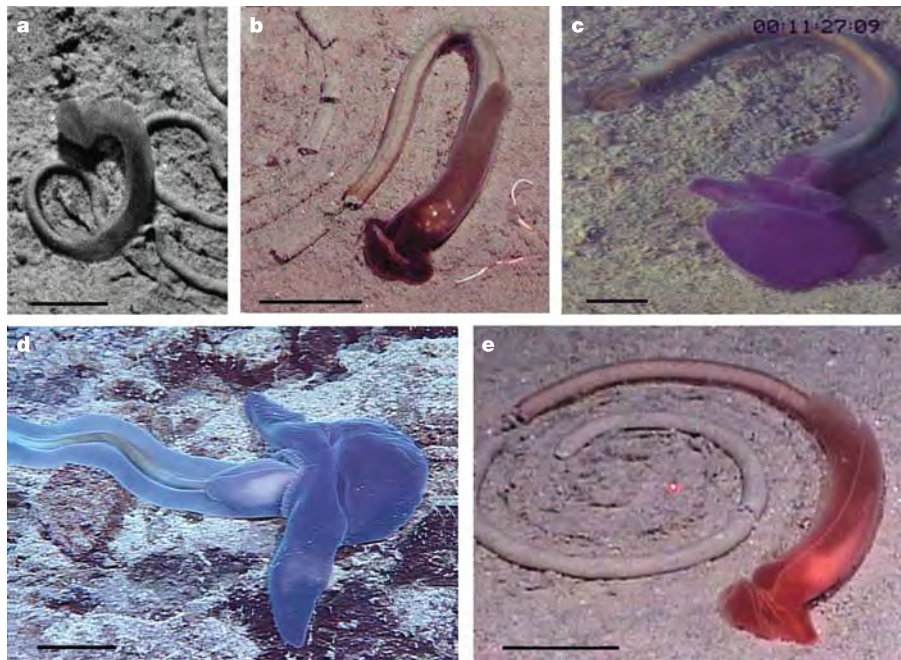


Figure 3 Deep-sea photographs of broad-collared enteropneusts not yet collected and described (depths, longitudes and latitudes are given in Table 1). **a, b**, Western Pacific (**a**) and Eastern Pacific (**b**) forms, with low dome-shaped proboscis and moderately broad collar. **c**, North Atlantic form with high dome-shaped proboscis and moderately broad

collar extended posterodorsally by two conspicuous lobes. **d**, Mid-Pacific form with high dome-shaped proboscis and very broad collar. **e**, A second North Atlantic form with small shield-shaped proboscis and moderately broad collar. Scale bar, 1 cm (**a, c**); 5 cm (**b, d, e**).

Table 1 Photographs of broad-collared, deep-sea (>1,500-m) enteropneusts

Designation	Quality of photo	Depth (m)	Location
Enteropneust ¹	High	4,735	30°S 177°W
Organism ²	Low	3,786	15°N 58°W
Organism ²	Low	3,681	10°N 92°W
No designation ³	Low	2,907	70°S 125°W
Enteropneust ⁴	Low	2,018	31°S 114°E
Enteropneust ⁴	Low	6,725	15°S 175°W
Lophenteropneust ⁵	Low	8,259	6°S 152°E
Lophenteropneust ⁵	Low	8,254	12°S 166°E
Lophenteropneust ⁶	Medium	5,160	12°N 159°W
Enteropneust ⁷	Medium	5,099	14°N 127°W
Enteropneust ⁸	Low	2,040	21°S 167°E
Enteropneust (Fig. 3a)	High	1,670	46°S 174°E
Enteropneust (Fig. 3b)*	High	2,715	43°N 127°W
Enteropneust (Fig. 3c)	High	3,000	53°N 35°W
Enteropneust (Fig. 3d)†	High	3,036	19°N 156°W
Enteropneust (Fig. 3e)	High	2,355	53°N 35°W

The table describes previously published photographs and those in Fig. 3; it omits deep-living enteropneusts with collars approximately circular in cross-section (*Saxipendium coronatum*¹⁶, *Glandiceps abyssicola*²² and undescribed^{23–25}).

*Six similar-looking enteropneusts were photographed on different dates at 36°N 123°W at depths between 3,200 and 3,498 m.

†A similar-looking enteropneust was photographed at 34°N 121°W at a depth of 1,960 m.

pharynx contained four acoel flatworms, evidently commensals, each about 1 mm long, and the post-hepatic intestine contained two intact harpacticoid copepods, possibly also commensals.

From an evolutionary point of view, our most significant finding is that the conspicuous width of the collar of *T. bullocki* is not due to the presence of tentacles. This species is therefore certainly not a lophenteropneust in the sense intended by Lemche⁵. This raises the question as to whether any of the enteropneusts visible in deep-sea photographs are actually lophenteropneusts. We therefore examined all of the photographic evidence, both published and unpublished (Fig. 3, Table 1). We found that photographs of low to moderate quality are inadequate to demonstrate the presence or absence of collar tentacles. Conversely, none of the high-resolution photographs with the potential to resolve collar tentacles ever showed such structures. Additionally, one observer in a submersible, who obtained a close view (but no photographs) of deep-sea enteropneusts, saw no tentacles¹⁸. In sum, there is no evidence that lophenteropneusts exist as originally described. They can be removed from the list of provocative evolutionary intermediates (although the evolutionary model they engendered—namely enteropneusts giving rise to pterobranchs—might still turn out to have merit on other grounds¹⁴). Although the lophenteropneust hypothesis resulted from a misinterpretation of low-resolution photographs, it has been a valuable stimulus for research on deep-sea hemichordates^{19–21}. □

Received 24 November 2004; accepted 24 January 2005; doi:10.1038/nature03382.

- Bourne, D. W. & Heezen, B. C. A wandering enteropneust from the abyssal Pacific, and the distribution of 'spiral' tracks on the sea floor. *Science* **150**, 60–63 (1965).
- Ewing, M. & Davis, R. A. in *Deep Sea Photography* (ed. Hersey, J. B.) 259–294 (Johns Hopkins Press, Baltimore, Maryland, 1967).
- Jacobs, S. S., Bruchhausen, P. M. & Bauer, E. B. *Hydrographic Stations, Bottom Photographs, Current Measurements, and Nephelometer Profiles*. *Eltanin Reports, Cruises 32–36* (Lamont-Doherty Geological Observatory, Palisades, New York, 1970).
- Heezen, B. C. & Hollister, C. D. *The Face of the Deep* (Oxford Univ. Press, New York, 1971).
- Lemche, H., Hansen, B., Madsen, F. J., Tendal, O. S. & Wolff, T. Hadal life as analyzed from photographs. *Vidensk. Meddr. Dansk Naturh. Foren.* **139**, 262–336 (1976).
- Thiel, H. Structural aspects of the deep-sea benthos. *AMBIO Spec. Rep.* **6**, 25–31 (1979).
- Foell, E. J. & Pawson, D. L. Photographs of invertebrate megafauna from abyssal depths of the north-eastern equatorial Pacific Ocean. *Ohio J. Sci.* **86**, 61–68 (1986).
- Gaillard, C. Recent organism traces and ichnofacies on the deep-sea floor off New Caledonia, southwestern Pacific. *Palaios* **6**, 302–315 (1991).
- Bateson, W. The ancestry of the Chordata. *Q. J. Microsc. Sci.* **26**, 535–571 (1886).
- Jollie, M. The origin of chordates. *Acta Zool. Stockh.* **54**, 81–100 (1973).
- Dilly, P. N. The pterobranch *Rhabdopleura compacta*: its nervous system and phylogenetic position. *Symp. Zool. Soc. Lond.* **36**, 1–16 (1975).
- Jefferies, R. P. S. in *Major Events in Early Vertebrate Evolution* (ed. Ahlberg, P. E.) 40–66 (Taylor & Francis, London, 2001).
- Lowe, C. J. et al. Anteroposterior patterning in hemichordates and the origin of the chordate nervous system. *Cell* **113**, 853–865 (2003).
- Cameron, C. B., Garey, J. R. & Swalla, B. J. Evolution of the chordate body plan: new insights from phylogenetic analysis of deuterostome phyla. *Proc. Natl Acad. Sci. USA* **97**, 4469–4474 (2000).

- Woodwick, K. H. in *Taxonomic Atlas of the Benthic Fauna of the Santa Maria Basin and the Western Santa Barbara Channel* Vol. 14 (eds Blake, J. A., Scott, P. H. & Lissner, A.) 251–259 (Santa Barbara Museum of Natural History, Santa Barbara, 1993).
- Woodwick, K. H. & Sensenbaugh, T. *Saxipendium coronatum*, new genus, new species (Hemichordata: Enteropneusta): the unusual spaghetti worms of the Galápagos Rift hydrothermal vents. *Proc. Biol. Soc. Wash.* **98**, 351–365 (1985).
- Young, C. M. in *Ecosystems of the Deep Oceans* (ed. Tyler, P. A.) 381–426 (Elsevier, Amsterdam, 2003).
- Pawson, D. Deep-sea dreams: diary of a mad lophenteropneust watcher. *Deep-Sea Newsl.* **32**, 6–7 (2003).
- Tendal, O. S. What became of Lemche's lophenteropneust? *Deep-Sea Newsl.* **27**, 21–24 (1998).
- Tendal, O. S. Lemche's lophenteropneust widely known but still an enigma. *Deep-Sea Newsl.* **28**, 8 (1999).
- Gage, J. D. & Tyler, P. A. *Deep Sea Biology, a Natural History of Organisms at the Deep-Ocean Floor* (Cambridge Univ. Press, Cambridge, 1991).
- Spengel, J. W. *Fauna und Flora des Golfes von Neapel* Monograph 18, *Die Enteropneusten des Golfes von Neapel* (Friedländer, Berlin, 1893).
- Menzies, R. J., George, R. Y. & Rowe, G. T. *Abyssal Environment and Ecology of the World Oceans* (Wiley, New York, 1973).
- Thorndike, E. M., Gerard, R. D., Sullivan, L. G. & Paul, A. Z. in *The Ocean Floor* (eds Scrutton, R. A. & Talwani, M.) 255–275 (Wiley, Chichester, 1982).
- Bett, B. J. UK Atlantic margin environmental survey: introduction and overview of bathyal benthic ecology. *Cont. Shelf Res.* **21**, 917–956 (2001).

Acknowledgements We thank C. B. Cameron and R. P. S. Jefferies for valuable taxonomic advice, and B. J. Bett, J. Barry, R. Logeman, B. H. Robison, R. J. Singleton, R. C. Vrijenhoek and the crews and pilots of the research vessels *Western Flyer*, *Akademik Mstislav Keldysh* and *G.O. Sars*, the manned submersibles 'Mir1' and 'Mir2', and the ROVs 'Tiburon' and 'Bathysaurus' for animal collection, photography, and unpublished observations. This study was partly under the auspices of the MAR-ECO Project within the Census of Marine Life Program and was partly supported by the New Zealand Foundation for Research, Science and Technology.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to N.D.H. (nholland@ucsd.edu).

Ancient co-speciation of simian foamy viruses and primates

William M. Switzer¹, Marco Salemi², Vedapuri Shanmugam¹, Feng Gao³, Mian-er Cong¹, Carla Kuiken⁴, Vinod Bhullar¹, Brigitte E. Beer⁵, Dominique Vallet⁶, Annie Gautier-Hion⁶, Zena Tooze⁷, Francois Villinger⁸, Edward C. Holmes⁹ & Walid Heneine¹

¹HIV and Retrovirology Branch, Division of HIV/AIDS Prevention, National Center for HIV, STD, and TB Prevention, Centers for Disease Control and Prevention, 1600 Clifton Road, MS G-19, Atlanta, Georgia 30333, USA

²Department of Pathology, Immunology and Laboratory Medicine, University of Florida, 1600 SW Archer Road, Gainesville, Florida 32610, USA

³Human Vaccine Institute, Duke University, 0112E Research Park III Durham, North Carolina 27710, USA

⁴Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

⁵Southern Research Institute, 431 Aviation Way, Frederick, Maryland 21701, USA

⁶CNRS-Université Rennes I, Ethology, Evolution, Ecology, Station Biologique, 35380 Paimpont, France

⁷Cercopan, 4 Ishie Lane, c/o Housing Estate, PO Box 826, Calabar, Cross River State, Nigeria

⁸Department of Pathology, Winship Cancer Center, Emory University, 1365B Clifton Road, Atlanta, Georgia 30322, USA

⁹Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

Although parasite–host co-speciation is a long-held hypothesis, convincing evidence for long-term co-speciation remains elusive, largely because of small numbers of hosts and parasites studied and uncertainty over rates of evolutionary change^{1–5}. Co-speciation is especially rare in RNA viruses, in which cross-species transfer is the dominant mode of evolution^{6–9}. Simian foamy viruses (SFVs) are ubiquitous, non-pathogenic retroviruses that infect all primates^{10,11}. Here we test the co-speciation

hypothesis in SFVs and their primate hosts by comparing the phylogenies of SFV polymerase and mitochondrial cytochrome oxidase subunit II from African and Asian monkeys and apes. The phylogenetic trees were remarkably congruent in both branching order and divergence times, strongly supporting co-speciation. Molecular clock calibrations revealed an extremely low rate of SFV evolution, 1.7×10^{-8} substitutions per site per year, making it the slowest-evolving RNA virus documented so far. These results indicate that SFVs might have co-speciated with Old World primates for at least 30 million years, making them the oldest known vertebrate RNA viruses.

Co-speciation of a host and its parasite is most likely if the parasite is transmitted vertically, causes a persistent infection and is present in a wide range of hosts⁹. Foamy viruses (FVs) are exogenous, persistent, and non-pathogenic retroviruses in the subfamily *Spumaretrovirinae* that have been isolated from a broad range of mammals, including horses, cows, cats and nonhuman primates (NHPs)^{10,11}.

We first inferred phylogenetic trees for the SFV polymerase gene (*pol*) from 44 different monkey and ape species originating from both Asia and Africa, and one New World monkey (Supplementary Information). The *pol* gene is highly conserved and is easily amplifiable from divergent SFVs¹¹. To infer the host phylogeny, we used mitochondrial (mtDNA) cytochrome oxidase subunit II (COII) sequences from 55 primate species (Supplementary Information). COII has previously been shown to be a powerful marker of primate phylogeny^{12–14}.

As expected, phylogenetic analysis of the COII sequences produced a tree similar to the known taxonomic relationships of the

primates, and to those estimated from other genes^{12–19}. The COII tree was composed of four monophyletic lineages within the Hominoidea (*Pan/Homo*, *Gorilla*, *Pongo* and *Hylobates*) and seven monophyletic lineages within the Cercopithecoidea, infraorder Catarrhini (Colobidae, *Trachypithecus*/*Pygathrix*/*Colobus*, *Cercopithecus*/*Chlorocebus*/*Erythrocebus*, *Allenopithecus*, *Macaca*, *Mandrillus*/*Cercocebus* and *Papio*/*Theropithecus*/*Lophocebus*) with the tree rooted with the New World *Ateles* sequence (Fig. 1a). Remarkably, the SFV tree formed two major monophyletic groups consisting of Hominoidea and Cercopithecoidea taxa similar to the clustering order of the COII tree (Fig. 1b). In addition, within the Hominoidea and Cercopithecoidea SFV clades, branching orders similar to those in the COII tree were observed (Fig. 1b). Within the Cercopithecoidea, the *Trachypithecus* SFV sequence was the basal taxon with the formation of seven additional clades, comprising *Cercopithecus*/*Chlorocebus*/*Erythrocebus*/African *Macaca*, *Allenopithecus*, *Cercocebus*/*Mandrillus*, African *Colobus*, Asian *Macaca*, *Papio*/*Theropithecus* and *Lophocebus* taxa. The SFV sequence from *Ateles* was excluded from the phylogenetic analysis because it did not seem to be a reliable outgroup. Similar trees were obtained by using both maximum-likelihood and neighbour-joining methods (data not shown), illustrating the strength of the primate and SFV phylogenetic relationships.

In total, we found statistically significant support ($P = 0.007$) for 22 co-speciation events between the 44 SFV and 55 COII-species specific sequences. Comparison of the SFV and COII trees (Fig. 2) by using reconciliation analyses shows both the congruent Hominoidea and Cercopithecoidea host superfamilies and SFV lineages and the identical clustering of specific subfamilies, such

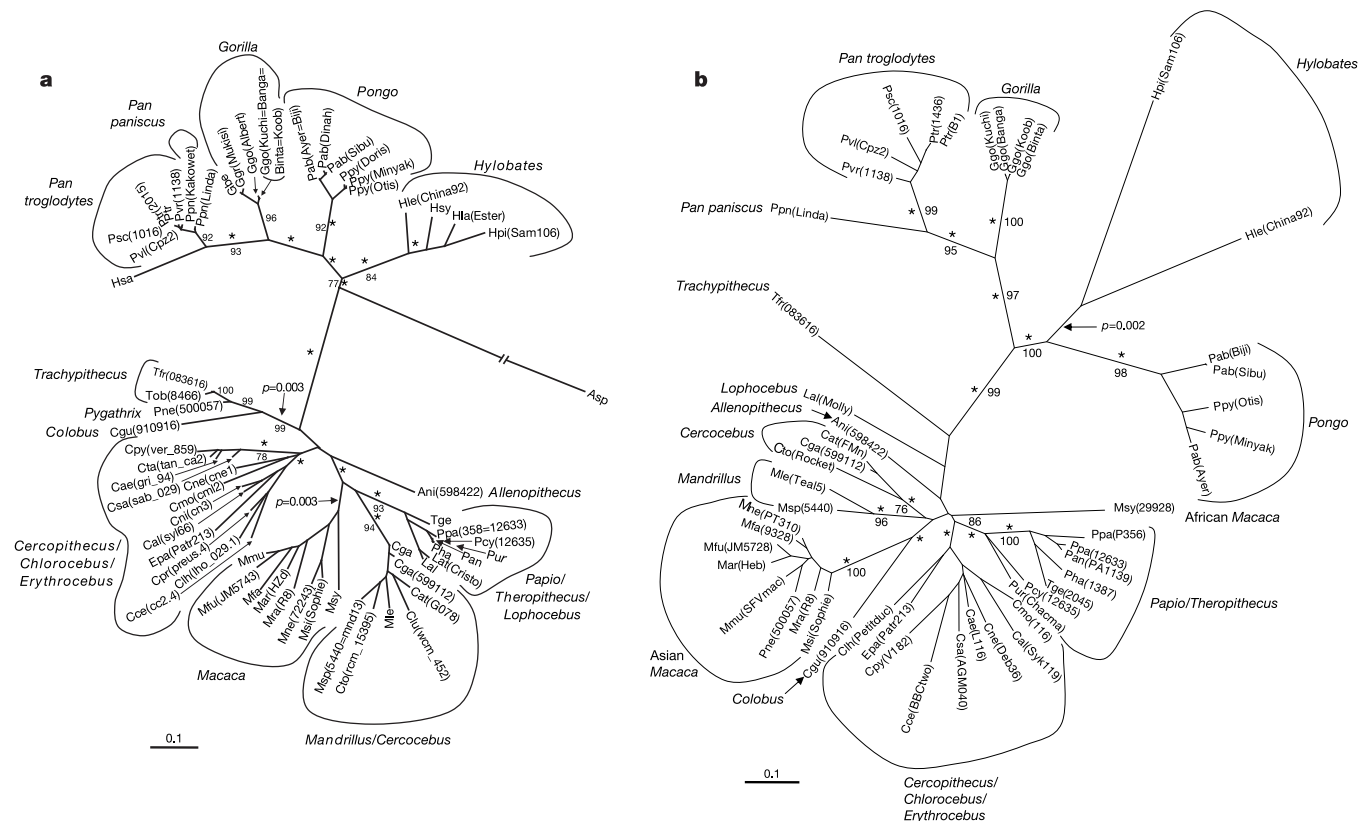


Figure 1 Congruence of host/parasite phylogenies. Phylogenetic relationships of (a) primate COII (500 base pairs) and (b) SFV *pol* sequences (425 base pairs) inferred by maximum-likelihood analysis. Support for the branching order was determined by 1,000 bootstrap replicates run on neighbour-joining trees using the maximum-likelihood substitution model and the zero branch length test (significance is indicated with P values

on the branches; asterisks indicate $P < 0.001$). Only bootstrap values 70% or greater are shown. Branch lengths are drawn to scale except for the *Ateles* (Asp) COII sequence; the bar indicates 0.1 nucleotide substitutions per site. Primate taxon codes are provided in Supplementary Information.

as Homininae (*Gorilla* and *Pan*), Ponginae (*Pongo*) and Cercopitheciinae (*Papio*). This is the first study to demonstrate such a comprehensive co-speciation relationship between exogenous viruses and their hosts.

Evidence for co-speciation of SFV and their primate hosts was also apparent on closer examination of sequences from individual primate species and subspecies within each clade (Fig. 1b). For example, SFV sequences from each of the Bornean (*Pongo pygmaeus*, Ppy) and Sumatran orangutans (*P. abelii*, Pab) formed well-supported sister taxa within the orangutan SFV lineage. However, one SFV sequence (PabAyer) from a Sumatran orangutan clustered with the Bornean SFVs (PpyOtis and PpyMinyak), indicating a possible interspecies transmission to this particular animal. Furthermore, SFV sequences from gibbons (*Hylobates*; HpiSam106 and HleChina92), gorillas (*Gorilla*; GgoBanga, GgoKuchi, GgoBinta and GgoKoob) and each chimpanzee species (*Pan paniscus* and

P. troglodytes; PpnLinda and PtrB1, PvlCpz2, Psc1016, Pvr1138, respectively) and from each subspecies of common chimpanzee (*P. t. troglodytes* (PtrB1), *P. t. verus* (Pvr1138), *P. t. vellerosus* (PvlCpz2) and *P. t. schweinfurthii* (Psc1016)) formed separate branches in the respective monophyletic ape lineages.

Although there was strong phylogenetic support for a broad co-evolution of primates and SFV, some instances of cross-species infections were evident. In addition to the SFV from a Sumatran orangutan noted above, the SFV sequence from a Douc langur (*Pygathrix nemaeus*, Pne500057) clustered with those from macaques and the African colobus sequence (*Colobus guereza*, Cgu910916) clustered with those from Asian macaques in the SFV tree (bootstrap support less than 60%) (Fig. 1b) rather than with SFV from an Asian langur (*Trachypithecus francoisi*, Tfr083616). Because all three animals were not wild-born, it is likely that their atypical SFVs were acquired from other primate species during captivity. Therefore, despite long-standing co-speciation with SFV, some primates may remain susceptible to infection with variant SFV from different species. In contrast to the monophyly of the *Papio* and *Lophocebus* taxa in the COII tree, the *Lophocebus* SFV sequence (LalMolly) formed a weakly supported lineage separate from the *Papio* SFV indicative of another cross-species infection. However, when SFVs from additional *Lophocebus* taxa were included in the analysis, they clustered with *Papio* SFV as expected (data not shown)^{16,19}.

A significant match between host and parasite phylogenies might also be produced under cross-species transmission if viruses are more likely to jump between related host species^{6–9}. To address this hypothesis we asked whether the divergence times of the COII and SFV trees are internally consistent. We found a strong linear relationship ($r = 0.8486$, $P < 0.0001$) between the branch lengths for the host and SFV gene trees, indicating that the accumulation of genetic diversity has occurred over an equivalent period in both data sets (Fig. 3). In addition, we found a strong agreement between the internode divergence times of the SFV and COII trees and the fossil record (Table 1). For example, we infer the separation of the chimpanzee and gorilla lineages to be 9.8–10.8 million years (Myr) ago with the SFV tree and 10.6–16.4 Myr ago with the COII tree, in general agreement with the divergence times of 7.4–8.8 Myr ago obtained by others^{18,19}. In addition, the separation of the chimpanzee and human lineages in the COII tree is in excellent agreement with the fossil record estimate of about 5–6 Myr ago^{18,19}. Given that SFV has a wide distribution among prosimian and simian primates, our data also imply that SFVs have coexisted in primates for an evolutionary history stretching at least 65 Myr. Furthermore, because divergent FVs asymptotically infect other mammalian orders¹⁰, the origin of FVs might even date back to the diversification of the placental mammals at more than 100 Myr

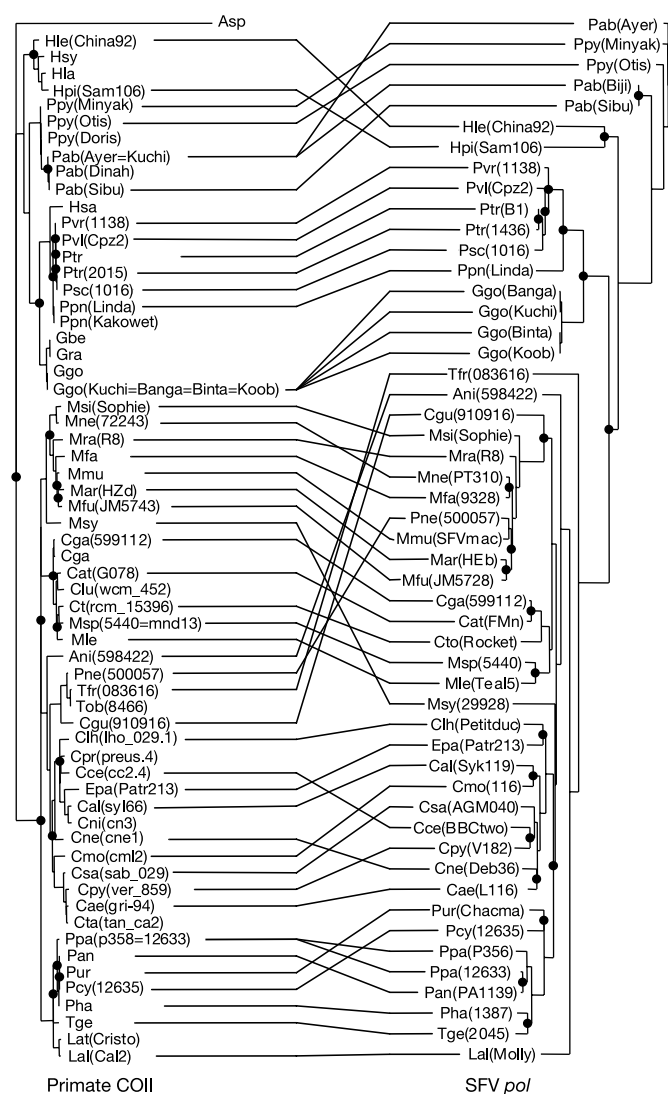


Figure 2 Co-evolutionary relationships of primate COII and SFV *pol* phylogenetic trees based on reconciliation analysis. Primate COII and SFV associations are indicated by connecting lines. Crossed lines indicate incongruent evolutionary host–virus relationships. SFV-infected hosts with identical COII sequences are indicated with multifurcated connecting lines. Hosts without SFV infection are indicated by the absence of a connecting line. Solid circles at nodes are inferred co-speciation events. The 22 co-speciation events were significantly higher than expected by chance in randomizations of both trees ($P = 0.007$).

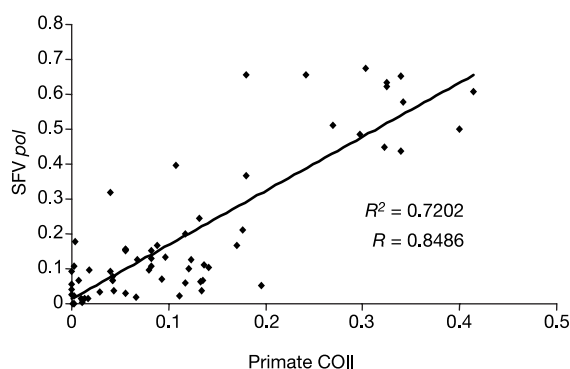


Figure 3 Correlation of branch lengths (substitutions per site) of primate COII and SFV *pol* phylogenetic trees. A statistically significant correlation ($P < 0.0001$) of the branch lengths was found.

Table 1 Estimates of branching times (Myr ago) within the Catarrhini, using host and parasite phylogenies

Branch point	SFV <i>pol</i>	COII	Fossil estimate*
Split Cercopithecoidea/Hominoidea	25–30	25–30	25–30
Cercopithecoidea	22.9–24.4	23.6–27.7	—
Hominoidea (=split Homonidae/Hylobatidae)	12–16	26.1	—
MRCA Hylobatidae	10.1–10.8	8.3–15.1	—
MRCA Ponginae	4.7–5.5	1.4–10.7	—
Split Homininae/Ponginae	n.a.†	18.5–23.5	~14
Split <i>Pan</i> / <i>Gorilla</i>	9.8–10.8	10.6–16.4‡	—
Split <i>Pan</i> / <i>Homo</i>	n.a.§	4.7–8.9	~5–6
MRCA <i>Pan</i>	5.9–6.6	1.2	—
Split <i>Cercocebus torquatus</i> / <i>M. sphinx</i>	—	1.4	—

Values for SFV and COII represent 95% confidence intervals; n.a., not available; MRCA, most recent common ancestor. The split of the Cercopithecoidea from the Hominoidea 25–30 Myr ago was used as a calibration point for the molecular dating of both SFV and COII sequences. Estimates and confidence intervals of the SFV and COII divergence times were determined on the maximum-likelihood tree topology by using the nonparametric rate smoothing algorithm implemented in the *r8s* program²¹. Values given without a range are the upper limit estimates.

*The fossil estimates for the Catarrhini are from refs 18 and 19.

†Homininae and Ponginae do not form a monophyletic clade in the SFV tree.

‡Includes *Homo sapiens* sequence.

§A human-specific foamy virus has yet to be identified.

ago²⁰. Thus, we propose FV to be the oldest known exogenous vertebrate RNA virus.

Under the assumption of co-speciation we were able to estimate the rate of nucleotide substitution in SFV. Although the presence of a strict molecular clock was weakly rejected for both the COII ($P = 0.005$) and SFV ($P = 0.004$) trees, the inferred substitution rates for the host and SFV genes were strikingly similar; assuming a calibration point of 25–30 Myr ago as the separation of the Hominoidea and Cercopithecoidea, the substitution rates for COII and SFV were estimated at $(1.16 \pm 0.35) \times 10^{-8}$ (MLE $\pm$ SE) and $(1.7 \pm 0.45) \times 10^{-8}$ substitutions per site per year, respectively. Similar substitution rates— $(1.03 \pm 0.42) \times 10^{-8}$ and $(1.83 \pm 0.93) \times 10^{-8}$, respectively—were obtained with the nonparametric rate smoothing program that does not assume a molecular clock²¹. SFV is therefore evolving so slowly that it approaches the rates recorded in protein-coding regions of mitochondrial DNA²². This substitution rate is the lowest ever reported for RNA viruses and is closer to those of DNA viruses (10^{-7} to 10^{-8} substitutions per site per year) and endogenous retroviruses ($(2.3\text{--}5.0) \times 10^{-9}$ substitutions per site per year) than to other exogenous RNA viruses such as HIV and influenza A virus (normally in the range 10^{-3} to 10^{-4} substitutions per site per year)^{9,23–25}.

The low substitution rates observed for SFV might be a function of their low rates of active replication, with most tissues latently infected¹⁰. FV expression is inhibited by negative regulation of an internal promoter unique to this group of retroviruses²⁶. In addition, FV regulatory proteins have also been shown to induce the expression of cellular genes, including that encoding interferon, that are known to reduce viral expression²⁶. Strong purifying selection and replication by means of a reverse transcriptase with a high fidelity will also reduce substitution rates. Whereas we found a greater number of synonymous (d_S) than non-synonymous (d_N) substitutions per site in the SFV *pol* sequences ($d_N/d_S = 0.161$), revealing the action of purifying selection, similar or greater levels of selective constraint have been observed in rapidly evolving RNA viruses²⁷ and in the primate COII sequences ($d_N/d_S = 0.069$). Although the fidelity of the SFV reverse transcriptase is unknown, our findings do not predict a high error rate per replication cycle as observed in HIV²⁵. However, it is unlikely that the accuracy of SFV reverse transcription alone can fully explain the observed low substitution rate, and thus lower SFV expression in persistently infected hosts must also have a major role in producing the low rate of evolutionary change observed here.

Although most NHP species investigated so far harbour distinct and species-specific SFV lineages, evidence supporting the existence of a human-specific FV is not yet available. Instead, all FVs

identified in humans are of simian origin in persons exposed to NHPs^{10,11}. Despite humans having a common evolution with NHPs and being susceptible to SFV infection, it is not understood why humans are not endemically infected with a distinct FV. It is possible that in the past humans were infected with an FV that became extinct because of reduced transmissibility, a finding consistent with the difficulty of human-to-human spread of SFV from accidentally infected humans^{10,11}. Alternatively, foci of endemically infected humans might exist but they have not yet been identified because of limited screening of small numbers of human populations¹⁰.

Thus, we provide extensive evidence supporting the co-speciation of SFV and primates. The non-pathogenic nature of SFV in persistently infected natural hosts, the ubiquity of SFV in primates, the congruent SFV and primate phylogenies and the similarity in divergence times all support an ancient history of host–parasite co-speciation. □

Methods

Primate specimens and SFV serology

Fresh blood specimens were obtained opportunistically from captive and wild-born primates in accordance with the animal care and use committees at each institution. For PCR analysis, nucleic acids were prepared from peripheral blood lymphocytes (PBLs) as described previously¹¹. All primates included in this study were determined as seropositive for SFV Gag antibodies in a combined antigen western blot assay as described¹¹. The primate taxonomic nomenclature used here was as described²⁸ and were coded using the first letter of the genus and the first two letters of the species or subspecies names with their house names or codes within parentheses (Supplementary Information).

SFV isolation and propagation

SFV isolates from selected NHP species (Supplementary Information) were obtained by co-cultivation of NHP PBLs with C2Th cells as reported previously¹¹.

Amplification of SFV and mtDNA COII DNA sequences

A total of 32 new SFV and 50 new primate mtDNA COII sequences were determined in this study. Polymerase chain reaction (PCR) was performed with 1 μ g of PBLs or PBL co-culture DNA in a generic nested amplification of a 465-base-pair SFV *pol* sequence by using methods described previously¹¹; 556-base-pair COII sequences were amplified by using standard PCR conditions with a 50 °C annealing temperature, 200 ng of PBL DNA, and the generic primers PCO2F2 (5'-A(C/T)GC(C/T)CTI(C/T)(C/T)CI(C/T)(A/G)ACA CTCACAACAAA-3') and PCO2R1 (5'-(G/A)AATAC(G/A)GG(T/C)CC(T/C)ATTTC (A/G)AAGATTTTTA-3'). The primers BCO2F2 (5'-ACGCCCTATCCTCAACACTCA CAACAAA-3') and BCO2R1 (5'-GAACACAGGTCTCTATTCAAGATTTTTA-3') or ACO2F2 (5'-ACGCCCTTTCTCTAACAACACTCAACAAA-3') and ACO2R1 (5'-GAATACGGG(C/T)CC(C/T)ATTTC(A/G)AAGATTTTTA-3') were used to amplify COII sequences from baboons or chimpanzees, respectively, using the same PCR conditions as above. COII sequences (1218 base pairs) were also amplified by reverse transcriptase PCR from mRNA obtained from the PBLs of selected primates (Supplementary Information) with the primers COII (5'-GCCACATCCCCTGTGCATAGAAGA-3') and APT6 (5'-GGTTTATAGAAAGTTGGTGGTGG-3') and PCR conditions 94 °C for 1 min, 55 °C for 30 s, 72 °C for 1.5 min, for a total of 30 cycles. Standard precautions were always taken to avoid PCR contamination. PCR products were revealed on 1.5% agarose gels stained with ethidium bromide, purified with a Qiaquick PCR purification kit (Qiagen) and sequenced in both directions with a BigDye terminator cycle kit and automated sequencers (Applied Biosystems).

Sequence analysis

SFV and COII sequences were aligned by using the ClustalW program (<http://www.cmbi.kun.nl/bioinf/CLUSTALW/>). To minimize the effects of nuclear mtDNA pseudogenes on our analyses, only COII sequences with open reading frames were used. Identical COII sequences from the same species were omitted. Little substitution saturation was observed in the SFV and COII sequences ($P < 0.00001$), as determined with the DAMBE program (<http://web.hku.hk/~xia/software/software.htm>), and they were therefore satisfactory for use in phylogenetic analyses. The best-fit phylogenetic model was determined with a likelihood ratio test and assumed the following nucleotide substitution rate pattern with maximum-likelihood estimated parameters: A to G and C to T = 12.3, A to C = 5.8, A to T and C to G = 3.1, G to T = 1.0; gamma-distributed rates (discrete approximation, eight categories) across sites with a maximum-likelihood estimate (MLE) of $\alpha = 0.37$, and unequal empirical nucleotide frequencies for the SFV alignment. This model was used to infer distance-based and maximum-likelihood trees. The best-fit substitution model for the COII sequences was HKY85 with gamma-distributed rates (transition/transversion = 12.1, $\alpha = 0.27$, MLE). This model was used to infer the COII maximum-likelihood tree. Neighbour-joining COII trees were inferred with LogDet/paralinear distances and the Moment Estimator technique enforcing the constraints of the HKY85 model. One thousand bootstrap replicates were used to assess the robustness of the final tree topology. All phylogenetic analyses were undertaken using the PAUP* package (<http://www.paup.csit.fsu.edu/>). To determine whether the SFV and COII sequences were behaving in a clock-like manner, the molecular-clock and variable-

rate phylogenetic trees were compared by using a likelihood ratio test.

Estimates and confidence intervals of the SFV divergence times were determined on the maximum-likelihood tree topology with the nonparametric rate smoothing algorithm implemented in the r8s program²¹. The branch lengths of the COII tree were re-estimated by enforcing a molecular clock with the PAUP* program. Confidence intervals for the COII divergence times were estimated with the r8s program by using the penalized likelihood algorithm²⁹ and 95% confidence intervals are given by 1.96σ . The split of the Cercopithecoidea from the Hominoidea 25–30 Myr ago was used as a calibration point for the molecular dating of both SFV and COII sequences^{18,19}. The number of synonymous (d_s) and non-synonymous (d_n) substitutions per site were estimated with the program Diverge in the Genetic Computer Group Wisconsin Package (http://www.accelrys.com/products/gcg_wisconsin_package).

Reconciliation analysis and comparison of branch lengths of the SFV and COII trees were performed with the TreeMap (v1.0) program in accordance with the author's instructions (<http://taxonomy.zoology.gla.ac.uk/rod/treemap.html>)³⁰. A single optimal reconstruction was found with the heuristic search option. The significance of the observed fit between the SFV and primate trees and branch lengths was determined by comparison with the distribution of the same measure of fit for 10,000 randomly generated trees or branch lengths by using the proportional-to-distinguishable model of the randomization test incorporated in TreeMap.

Received 7 September; accepted 22 December 2004; doi:10.1038/nature03341.

- Hafner, M. S. *et al.* Disparate rates of molecular evolution in cospeciating hosts and parasites. *Science* **265**, 1087–1090 (1994).
- Page, R. D. M. Temporal congruence and cladistic analysis of biogeography and cospeciation. *Syst. Zool.* **39**, 205–226 (1990).
- Brooks, D. R. Analysis of host-parasite coevolution. *Int. J. Parasitol.* **17**, 291–297 (1987).
- Charrel, R. N., De Micco, P. & de Lamballerie, X. Phylogenetic analysis of GB viruses A and C: evidence for cospeciation between virus isolates and their primate hosts. *J. Gen. Virol.* **80**, 2329–2335 (1999).
- McGeoch, D. J., Dolan, A. & Ralph, A. C. Toward a comprehensive phylogeny for mammalian and avian herpesviruses. *J. Virol.* **74**, 10401–10406 (2000).
- Charleston, M. A. & Robertson, D. L. Preferential host switching by primate lentiviruses can account for phylogenetic similarity with the primate phylogeny. *Syst. Biol.* **51**, 528–535 (2002).
- Salemi, M. *et al.* Mosaic genomes of the six major primate lentivirus lineages revealed by phylogenetic analyses. *J. Virol.* **77**, 7202–7213 (2003).
- Jackson, A. P. & Charleston, M. A. A cophylogenetic perspective of RNA-virus evolution. *Mol. Biol. Evol.* **21**, 45–57 (2004).
- Holmes, E. C. Molecular clocks and the puzzle of RNA virus origins. *J. Virol.* **77**, 3893–3897 (2003).
- Meiering, C. D. & Linial, M. L. Historical perspective of foamy virus epidemiology and infection. *Clin. Microbiol. Rev.* **14**, 165–176 (2001).
- Hussain, A. I. *et al.* Screening for simian foamy virus infection by using a combined antigen Western blot assay: evidence for a wide distribution among Old World primates and identification of four new divergent viruses. *Virology* **309**, 248–257 (2003).
- Disotell, T. R., Honeycutt, R. L. & Ruvofo, M. Mitochondrial DNA phylogeny of the Old-World monkey tribe Papionini. *Mol. Biol. Evol.* **9**, 1–13 (1992).
- Ruvofo, M., Disotell, T. R., Allard, M. W., Brown, W. M. & Honeycutt, R. L. Resolution of the African hominoid trichotomy by use of a mitochondrial gene sequence. *Proc. Natl Acad. Sci. USA* **88**, 1570–1574 (1991).
- Adkins, R. M. & Honeycutt, R. L. Evolution of the primate cytochrome C oxidase subunit II gene. *J. Mol. Evol.* **38**, 215–231 (1994).
- van der Kuyl, A. C., Kuiken, C. L., Dekker, J. T. & Goudsmit, J. Phylogeny of African monkeys based upon mitochondrial 12S rRNA sequences. *J. Mol. Evol.* **40**, 173–180 (1995).
- Page, S. L., Chiu, C. H. & Goodman, M. Molecular phylogeny of Old World monkeys (*Cercopithecoidea*) as inferred from gamma-globin DNA sequences. *Mol. Phylogenet. Evol.* **13**, 348–359 (1999).
- Harris, E. E. & Disotell, T. R. Nuclear gene trees and the phylogenetic relationships of the mangabeys (Primates: Papionini). *Mol. Biol. Evol.* **15**, 892–900 (1998).
- Goodman, M. *et al.* Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence. *Mol. Phylogenet. Evol.* **9**, 585–598 (1998).
- Page, S. L. & Goodman, M. Catarrhine phylogeny: noncoding DNA evidence for a diphyletic origin of the mangabeys and for a human-chimpanzee clade. *Mol. Phylogenet. Evol.* **18**, 14–25 (2001).
- Murphy, W. J. *et al.* Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294**, 2348–2351 (2001).
- Sanderson, M. J. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**, 301–302 (2003).
- Ruvofo, M. *et al.* Mitochondrial COII sequences and modern human origins. *Mol. Biol. Evol.* **10**, 1115–1135 (1993).
- Shadan, F. F. & Villarreal, L. P. Coevolution of persistently infecting small DNA viruses and their hosts linked to host-interactive regulatory domains. *Proc. Natl Acad. Sci. USA* **90**, 4117–4121 (1993).
- Johnson, W. E. & Coffin, J. M. Constructing primate phylogenies from ancient retrovirus sequences. *Proc. Natl Acad. Sci. USA* **96**, 10254–10260 (1999).
- Suzuki, Y., Yamaguchi-Kabata, Y. & Gojobori, T. Nucleotide substitution rates of HIV-1. *AIDS Rev.* **2**, 39–47 (2000).
- Löschelt, M. Foamy virus transactivation and gene expression. *Curr. Top. Microbiol. Immunol.* **277**, 27–61 (2003).
- Woelck, C. H. & Holmes, E. C. Reduced positive selection in vector-borne RNA viruses. *Mol. Biol. Evol.* **19**, 2333–2336 (2002).
- Groves, C. *Primate Taxonomy* (Smithsonian Institution Press, Washington DC, 2001).
- Sanderson, M. J. Estimate absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol. Biol. Evol.* **19**, 101–109 (2002).
- Page, R. D. M. Parallel phylogenies: reconstructing the history of host-parasite assemblages. *Cladistics* **10**, 155–173 (1994).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank the veterinary staff at all zoological gardens and primate centres who kindly provided blood specimens from the primates living at their institutions; R. Heberle and P. Johnston for the SFV-infected baboon and orangutan isolates; A. Hussain and A. Wright for expert technical assistance; and A. Vandamme, C. Coulbaly, M. Peeters, F. Bibollet-Ruche, V. Hirsch, J. Allan, T. Butler and H. McClure for providing additional primate samples for this study.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to W.M.S. (bis3@cdc.gov). The GenBank accession numbers for the 32 new SFV *pol* and 50 new primate mtDNA COII sequences are AY686124–AY686148 and AY686150–AY686206.

Genetic relatedness predicts South African migrant workers' remittances to their families

S. Bowles¹ & D. Posel²

¹Santa Fe Institute, 1399 Hyde Park Rd, Santa Fe, New Mexico 87501, USA; and ²Università degli Studi di Siena, Piazza San Francesco, Siena 53100, Italy

²University of KwaZulu-Natal, King George V Avenue, Durban 4001, South Africa

Inclusive fitness models^{1,2} predict many commonly observed behaviours: among humans, studies of within-household violence³, the allocation of food^{4,5} and child care⁶ find that people favour those to whom they are more closely related. In some cases however, kin-altruism effects appear to be modest^{7–9}. Do individuals favour kin to the extent that kin-altruism models predict? Data on remittances sent by South African migrant workers to their rural households of origin allow an explicit test, to our knowledge the first of its kind for humans. Using estimates of the fitness benefits and costs associated with the remittance, the genetic relatedness of the migrant to the beneficiaries of the transfer, and their age- and sex-specific reproductive values, we estimate the level of remittance that maximizes the migrant worker's inclusive fitness. This is a much better predictor of observed remittances than is average relatedness, even when we take account (by means of a multiple regression) of covarying influences on the level of remittance. But the effect is modest: less than a third of the observed level of remittances can be explained by our kin-altruism model.

Migrants' remittances provide a rare window into the allocation of resources within a household, as intra-household transfers are typically not measured in surveys. The large and genetically heterogeneous nature of rural African households makes migrants and their households of origin an ideal database. The data for our study come from a nationally representative survey in 1993 of approximately 9,000 households¹⁰.

We have complete data on the income, remittances received, land ownership and composition of the household of origin, and on each migrant worker's age and schooling for 539 black male migrants. Virtually all migrants in the sample sent remittances in the year preceding the survey in 1993, and on average they sent almost half of their urban wage. Figure 1 shows the distribution of average relatedness of the migrant to the household of origin. (Consanguineous marriages are exceptionally rare in South Africa¹¹, so we assume that the migrant and his wife are unrelated.)

Hamilton's rule states that conferring a fitness benefit (b) by helping another at a cost of (c) to oneself will be selected for if $rb > c$, where r is the genetic relatedness of the donor to the beneficiary. Here we consider a case in which the choice is not

simply whether to help, but how much to help. The relationship between fitness and income will be increasing but concave because there are diminishing fitness returns to material resources. This is especially true at the very low levels of income prevalent among migrants and their households¹². As a result, the values of b and c above will vary with the amount of the remittance.

To express the concave relationship between material resources available to an individual and fitness in a way that yields a closed-form prediction of the fitness-maximizing remittance, we assume the relationship is logarithmic:

$$f^i = v \ln x \quad (1)$$

where f^i is the expected contribution to the gene pool in the future by individual i , v is the multiplicative baseline fitness and $x \geq 1$ is the income available to the individual in a given period. Income contributes to reproductive success by increasing the likelihood of survival to reproductive age, the expected period of reproductive life for those who attain this age, and the number of resulting offspring surviving to reproductive age. The inclusive fitness of individual m (the migrant) is:

$$F^m = \sum r^{im} f^i \quad (2)$$

where r^{im} is the relatedness of the migrant to individual i , one of the $n + 1$ persons with whom he interacts (the n members of the migrant's household of origin, plus himself, with $r^{mm} = 1$) and the summation is over $i = 1, \dots, n + 1$. The average relatedness of the migrant to his household members is $r = (\sum r^{im} - 1)/n$.

Let the pre-remittance income available to each household member be y and that available to the migrant (his wage) be w , so that after a remittance of size t , the migrant has $w - t$ and each household member $y + t/n$. We assume that the income and (for the moment) the remittance are shared equally among the household members. (In our empirical calculations of the income available to household members, we count children aged less than 6 years as half of an adult equivalent, but to avoid notational clutter in the explication of the model we ignore the presence of less than adult-equivalent children.)

Differentiating equation (2) with respect to t and setting the result equal to zero, we find that the F -maximizing transfer (t^*) equalizes the marginal fitness cost to the migrant (that is, $1/(w - t)$) and the marginal fitness benefit to the recipients, weighted by their average relatedness to the migrant (that is, $r/(y + t/n)$) or equivalently:

$$r = \frac{(y + t/n)}{(w - t)} \quad (3)$$

Condition (3) requires the optimal transfer (if one exists) to

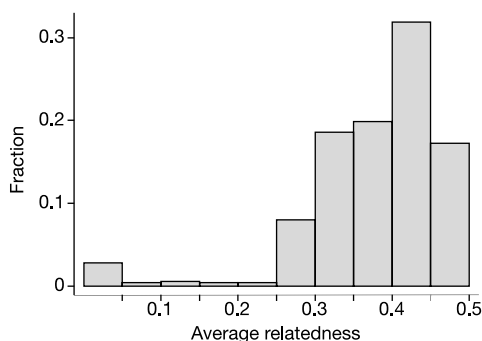


Figure 1 Average relatedness of migrants to their households of origin. More than three in ten of the migrants' households of origin include three generations, and one in five includes the migrant's aunt, uncle, niece, nephew or grandchild. The average degree of relatedness between the migrant and the members of the household is 0.37.

implement a ratio of the average post-transfer income of the household to that of the migrant (the right hand side of equation (3)) that is equal to the average relatedness of the household members to the migrant. This is accomplished (rearranging equation (3)) when:

$$t^* = \frac{(wr - y)}{(r + 1/n)} \text{ if } wr > y \text{ and otherwise } t^* = 0 \quad (4)$$

The numerator of the right-hand side of equation (4) is a version of Hamilton's rule: some positive transfer is F -maximizing if $r/y > 1/w$. Here $b (= 1/y)$ and $c (= 1/w)$ are the marginal fitness benefits and costs of an arbitrarily small transfer. The remittance that satisfies equation (4) is unique owing to the concavity of equation (1) and the resulting monotonicity of the marginal fitness costs and benefits functions. As can be seen from equation (4), the F -maximizing transfer (if one exists) increases with n ; in larger households the remittance is spread among more members, so diminishing marginal fitness returns are less pronounced. Figure 2 illustrates the optimal transfer.

The effect on inclusive fitness of assistance to a genetically related individual depends not only on the degree of relatedness but on the age, sex and other characteristics relevant to the reproductive value of the beneficiary^{13,14}. For example, owing to the very high levels of mortality among rural black children in South Africa^{12,15}, a non-negligible fraction of the children benefiting from remittances will not reach reproductive age, irrespective of the level of the remittance. Also, some closely related beneficiaries (parents, for example) have few or no years of reproductive activity remaining. Finally, fitness depends on the relative (not absolute) contribution to future gene pools¹⁶, so in a rapidly growing population such as South Africa's in the early 1990s earlier fecundity should be favoured over later, thus enhancing the reproductive value of those who have reached reproductive age compared to those who have not. An inclusive-fitness-maximizing migrant would thus give more if the members of his household of origin were his adolescent children rather than his infant children or his aged parents (despite the fact that they are equally related to the migrant). And he would also give more if he were near the end of his reproductively active years rather than at the beginning.

Let v^j be an age- and sex-specific effect such that the individual fitness of the age-sex class j is $f^j = v^j \ln x$. Then, using equations (1) and (2):

$$F^m = v^m \ln(w - t) + \sum v^j r^{im} \ln(y + t/n) \quad (5)$$

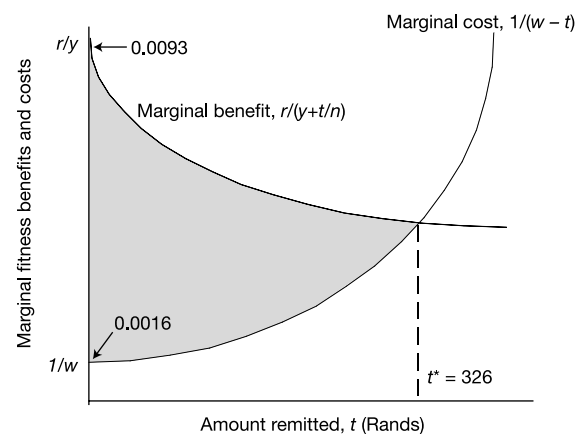


Figure 2 The optimal remittance t^* equates the marginal fitness cost to the migrant with the sum of the relatedness-weighted marginal fitness benefits to the members of the household of origin. The total inclusive fitness is the shaded area. The numerical values are for a migrant with mean values of r , y , n and w .

where v^m and v^i are the reproductive values (respectively) of the migrant and the i th household member ($i = 1 \dots n$). Now we define the reproductive-value-adjusted average relatedness of the household, $r^v = \sum v^i r^{im} / n$. Maximizing F^m as above, the resulting optimal transfer (t^v) equates the ratio of the two post-transfer incomes to r^v / v^m (the analogue of equation (3) above), or:

$$t^v = \frac{(wr^v - yv^m)}{(v^m/n + r^v)} \text{ if } wr^v > yv^m \text{ and otherwise } t^v = 0 \quad (6)$$

which is identical to equation (4) if $v^i = 1$, for all i .

We estimate the v^i in equations (5) and (6) as the present value of the sequence of years of reproductive life that an individual of age j may expect, discounted by the rate of population growth and normalized to equal one for a child aged 1 year. Estimates and methods of calculation of v^i appear in Fig. 3. The rise in reproductive value before attaining reproductive age is expected in societies with high child mortality and substantial rates of population growth, and is observed in other southern African data¹⁷.

We use the migrant's schooling and age to estimate his expected wage, which, along with data on the other variables above, allow an estimate of t^* and t^v for each migrant. The first numeric column of Table 1 gives the simple correlations ρ between observed transfers t on the one hand and our measures of relatedness and optimal transfers on the other. As can be seen, t^* and t^v are more highly correlated with t than are r and r^v , and taking account of reproductive values significantly increases the correlation of the observed and predicted remittances (from 0.25 to 0.32). Using the larger correlation, one-tenth (0.32²) of the variance of observed remittances is explained by the inclusive-fitness-maximizing transfer.

To estimate the causal importance of inclusive fitness in explaining migrants' remittances we need to control for other influences on t that do not represent inclusive-fitness effects but that are correlated with t^v . The age and years of schooling of the migrant, for example, are a measure of his economic standing that may be associated with claims on his income independently of genetic relatedness. Older, better-educated men are likely to maintain larger social networks and may be expected to give more. Conversely, the economic need of the household of origin may influence

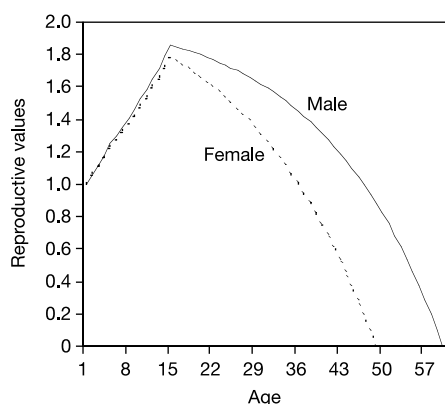


Figure 3 Normalized reproductive values: rural black South Africans around 1993 (refs 15, 21). Let d^a be the age- and sex-specific probability of not surviving from age a to $a + 1$ and let m^q take a value of 1 if q is one of the reproductive years (15 to T where T is the age at the end of one's reproductive years: 49 years for women, 60 years for men) and zero otherwise. We define the present expected value of the q th year of life of someone who is now in their j th year, discounted at the rate of population growth g , as $V^q = m^q \Pi [(1 - d^a)/(1 + g)]$, where the product is over $a = j, \dots, q$. Then the reproductive value of a person j years old is $V^j = \sum V^q$, where the summation is over $q = j, \dots, T$. Reproductive values are normalized using sex-specific reproductive values at year 1 so $v^j = V^j / V^1$.

remittance behaviour independently of its effect on the marginal fitness benefits of additional income. We therefore predict t (by ordinary least-squares methods) using the migrant's age, schooling and land ownership of the household and presence of the migrant's wife in the household as explanatory variables in addition to t^v .

When conditioned on the above controls, the estimated normalized regression coefficient of t^v (β , shown in the second numeric column of Table 1) indicates that a standard deviation difference in t^v is associated with a difference in predicted remittances of 18 per cent of a standard deviation, controlling for the effects of these other variables. To see what this estimate means, consider a migrant whose relatedness to his household of origin makes the optimal remittance zero ($wr^v - yv^m = 0$, so $t^v = 0$) but with sample mean values for all of the other predictors. This hypothetical migrant would remit 23 per cent less than an otherwise identical migrant with the sample mean relatedness and hence also the sample mean t^v . Thus, although highly statistically significant, the difference in remittance behaviour accounted for by the inclusive-fitness model is modest in size. (The analogous estimate using t^* is 15 per cent.) The fact that (in results not shown here) remittances are slightly better predicted using reproductive values assuming zero population growth (which was approximately the case in the ancestral populations of the migrants studied here¹⁷) is consistent with the view that contemporary behaviour may be an adaptation to past conditions.

Have we underestimated the inclusive-fitness effect? In all of these estimated equations, households in which the migrant's wife is present receive more (by 45 per cent of the observed mean remittance ($p < 0.001$) in the equation using t^v). The wife's presence in the household may increase remittances both for inclusive-fitness reasons (for example, care of the migrant's children, enhanced future reproductive success with his wife) and as a result of altruistic motives towards non-kin of the type documented in recent behavioural experiments¹⁸.

To incorporate the effect of the wife's care of existing offspring in our estimate of the optimal transfer, we redefine t^v assuming that if both the migrant's wife and any of his children are resident in the household, the children and the mother receive (and share equally) all of the remittance. To take account of the wife's contribution to the migrant's prospective future offspring we add one or more fictive children of the migrant to the household in calculating r^v , each with a reproductive value of less than that of a newborn infant because of the passage of time between the present and its birth some years hence (should the mother survive until then). The number of such hypothetical additions is the expected total number of children born to a rural black woman during her reproductive years¹⁹, minus the number of the migrant's children currently resident in the household, conditional on the wife having sufficient remaining years of reproductive life for this to occur.

Taking account of the wife's contributions in these two ways, we

Table 1 The statistical relationship between observed and predicted remittances

Observed remittances, t	Simple correlation, ρ (p)	Normalized regression coefficient, β (p)
Average relatedness, r	-0.03 (0.470)	0.11 (0.004)
Reproductive-value-weighted relatedness, r^v	0.14 (0.001)	0.13 (0.001)
Predicted remittance, using r , t^*	0.25 (<0.001)	0.11 (0.064)
Predicted remittance using r^v , t^v	0.32 (<0.001)	0.18 (0.005)
Predicted remittance (wife effect), t^w	0.35 (<0.001)	0.21 (0.001)

ρ is the correlation of the indicated variable with observed remittances t (with p -values in parentheses). β is the normalized regression coefficient of the variable indicated in the ordinary least-squares equation in which, in addition to the variables indicated here, predictors included the migrant's age, schooling, the land owned by the household and whether the migrant's wife is present in the household, as well as place-of-origin controls. Thus, a standard deviation difference in t^w holding these other variables constant is associated with a fifth of a standard deviation difference (21%) in t .

estimate for each migrant a new F -maximizing remittance of t^w . Using this predictor in the multiple regression described above and in Table 1, and the same methods as above, inclusive fitness accounts for 29 per cent of the observed remittances. The 'wife present' effect remains large (38 per cent of the observed mean remittance ($p < 0.0001$)) in this equation, suggesting that although the wife's contribution to inclusive fitness (as we have modelled it) helps to explain remittances, most of the 'wife present' effect cannot be explained this way.

The measurement of y , w and r^{im} are all subject to error, and this imparts a downward bias to our estimates of β . Moreover, other functional forms would yield different predictions. Finally, in many cases migrants have formed households and fathered children in the locality of their work²⁰, thus raising the marginal fitness costs of remitting in ways that our model does not capture. We experimented by re-estimating t^v assuming plausible (but hypothetical) values for the composition of the typical migrant's secondary family. This reduces the mean predicted transfer, as expected, but it does not increase the fraction of the remittance accounted for by inclusive fitness. We doubt that addressing these limitations would alter the conclusion that inclusive fitness explains part of remittance behaviour, but not all of it. □

Received 15 November 2004; accepted 31 January 2005; doi:10.1038/nature03420.

- Hamilton, W. D. The genetical evolution of social behavior. *J. Theor. Biol.* **7**, 1–52 (1964).
- Taylor, P. Inclusive fitness arguments in genetic models of behavior. *J. Math. Biol.* **34**, 654–674 (1996).
- Daly, M. & Wilson, M. Homicide and kinship. *Am. Anthropol.* **84**, 372–378 (1982).
- Gurven, M., Hill, K. & Kaplan, H. From forest to reservation: Transitions in food sharing among the Ache of Paraguay. *J. Anthropol. Res.* **58**, 93–120 (2002).
- Case, A., Lin, I.-F. & McLanahan, S. How hungry is the selfish gene? *Econ. J.* **110**, 781–804 (2000).
- Anderson, K., Kaplan, H., Lam, D. & Lancaster, J. Paternal care by genetic fathers and stepfathers. II: Reports by Xhosa high school students. *Evol. Hum. Behav.* **20**, 433–451 (1999).
- Gurven, M., Hill, K., Kaplan, H., Hurtado, A. M. & Lyles, R. Food transfers among Hiwi foragers of Venezuela: Tests of reciprocity. *Hum. Ecol.* **28**, 171–218 (2000).
- Kaplan, H. & Hill, K. Food sharing among Ache foragers: tests of explanatory hypotheses. *Curr. Anthropol.* **26**, 223–246 (1985).
- Smith, E. A., Bird, R. B. & Bird, D. W. Risk and reciprocity in Meriam food-sharing. *Evol. Hum. Behav.* **23**, 297–321 (2002).
- South Africa Integrated Household Survey: The Project for Statistics on Living Standards and Development (A Living Standards Measurement Survey of the World Bank, South African Labour and Development Research Unit, Univ. Cape Town, Cape Town, 1993); (<http://www.worldbank.org/html/prdph/lms/country/za94/za94home.html>).
- Stevenson, A. C., Johnston, H. A., Stewart, M. I. & Golding, D. R. Congenital malformations: A report of a study of series of consecutive births in 24 centres. *Bull. World Health Org.* **34** (suppl.), 1–127 (1966).
- Mencarini, L. An Analysis of Fertility and Infant Mortality in South Africa based on 1993 LSMS Data 109–128 Pre-conf. Proc. of Third African Population Conf. 'African Population in the 21st Century' (Durban, South Africa, 6–10 December 1999) (UAPS, Dakar, Senegal, 1999).
- Fisher, R. A. *The Genetical Theory of Natural Selection* 27–30 (Clarendon, Oxford, 1930).
- Grafen, A. A Theory of Fisher's Reproductive Value (St John's College, Oxford, 2004).
- Kahn, K., Tollman, S., Garenne, M. & Gear, J. S. S. Who dies from what? Determining cause of death in South Africa's rural north-east. *Trop. Med. Int. Health* **4**, 433–441 (1999).
- Grafen, A. Formal darwinism, the individual-as-maximizing-agent analogy, and bet-hedging. *Proc. R. Soc.* **266**, 799–803 (1999).
- Pennington, R. & Harpending, H. *The Structure of an African Pastoralist Community: Demography, History, and Ecology of the Ngamiland Herero* (Clarendon, Oxford, 1993).
- Fehr, E. & Fischbacher, U. The nature of human altruism—proximate patterns and evolutionary origins. *Nature* **425**, 785–791 (2003).
- South Africa Demographic and Health Survey 1998 (Preliminary Report, Department of Health, 1998); (<http://www.doh.gov.za/facts/1998/sadhs98/>).
- Posel, D. Intra-family transfers and income-pooling: A study of remittances in KwaZulu-Natal. *S. Afr. J. Econ.* **69**, 501–528 (2001).
- World Health Organization. *The World Health Report, 1999: Making a Difference* 90, Table 1, Statistical Appendix; (www.who.int/entity/whr/1999/en) (WHO, Geneva, 1999).

Acknowledgements The authors contributed equally to this paper. D.P. thanks the Woodrow Wilson School at Princeton University for the post-doctoral leave supporting this research. S.B. thanks the Behavioural Sciences Program of the Santa Fe Institute for financial support. We have benefited from the contributions of R. Boyd, N. Borgerhoff Mulder, N. Folbre, H. Kaplan, R. Rowthorn and D. E. Smith.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.P. (posel@ukzn.ac.za).

Genetic effects on sperm design in the zebra finch

T. R. Birkhead¹, E. J. Pellatt¹, P. Brekke¹, R. Yeates¹ & H. Castillo-Juarez²

¹Department of Animal & Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK

²Universidad Autónoma Metropolitana-Xochimilco, Calzada del Hueso 1100, Distrito Federal, CP 04960, Mexico

Sperm design and function are important determinants of male reproductive success and are expected to be under strong selection^{1,2}. The way that spermatozoa phenotypes evolve is poorly understood, because there have been few studies of the quantitative genetics of sperm^{3–5}. Here we show, in the zebra finch *Taeniopygia guttata*, an extraordinary degree of inter-male variation in sperm design that is independent of sperm swimming velocity. A quantitative genetics study using data from over 900 zebra finches in a complex breeding experiment showed that sperm head, mid-piece and flagellum length are heritable, that negative genetic correlations exist between sperm traits, and that significant indirect (maternal) genetic effects exist. Selection on the zebra finch sperm phenotype may be low because sperm competition is infrequent in this species⁶, and this, in combination with negative genetic correlations and maternal genetic effects, may account for the variation in sperm phenotype between males. These results have important implications for the evolution of sperm in other taxa.

The primary function of the male gamete, the spermatozoon, is to fertilize ova. Sperm are expected to be under strong selection to be efficient fertilizers for two reasons. First, within the female reproductive tract sperm face numerous physical, chemical and immunological barriers that result in only a tiny subset of inseminated sperm reaching the ovum^{7,8}. Second, widespread promiscuity among females results in sperm competition between males, which favours males whose sperm are effective competitors¹. Interspecific differences in these evolutionary forces combined with phylogenetic effects⁹ probably account for the fact that despite their common purpose, sperm vary more dramatically in size and design across species than any other cell type¹⁰. As well as the marked interspecific differences in sperm design, considerable inter-male variation in sperm phenotype exists within species¹¹.

Here, we aimed to investigate the underlying causes of inter-male variation in sperm design in the zebra finch *Taeniopygia guttata* by examining the relationship between sperm phenotype, sperm swimming velocity and the quantitative genetics of sperm traits. Heritabilities and other genetic parameters were estimated using the animal model¹² from a combined full-sibling and half-sibling animal breeding design involving 81 sires producing two sons from each of six dams (972 male offspring; entire pedigree comprising 1,526 individuals) and analysed using the multiple-trait, derivative-free, restricted maximum-likelihood program (MTDFREML)¹³.

Within males the repeatability of sperm traits was substantial, particularly for mid-piece (0.75) and flagellum (0.84) length (repeatability for head length was 0.52, all degrees of freedom (d.f.) = 913, 3,656; $P < 0.001$), and we have shown elsewhere that within-male variation in sperm design is highly consistent across ejaculates and time¹⁴. In contrast, between-male phenotypic variation in sperm design was considerable (Fig. 1 and Table 1). The inter-male coefficients of variation in sperm flagellum length and mid-piece length were two and four times greater, respectively, than for the linear morphological trait of male tarsus length (coefficient of variation = 3.38, $n = 972$ males).

Sperm motility is determined largely by the flagellum and the

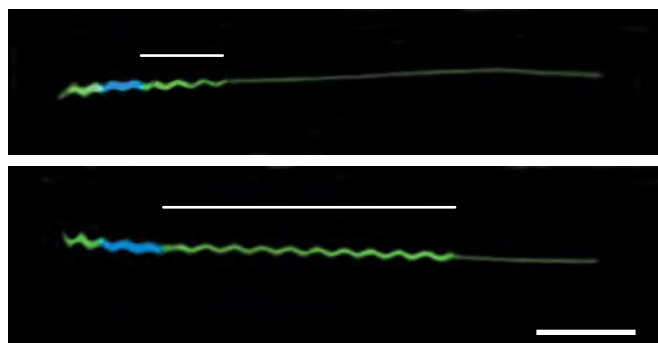


Figure 1 Examples of sperm of approximately the same total length from two zebra finch males showing the marked difference in size of the mid-piece (that is, the mitochondrial helix (bright green and indicated by a horizontal line)). The nucleus is stained blue and the acrosome at the tip is stained green. Scale bar, 10 μm .

mid-piece—flagellar forces are fuelled by ATP supplied by mitochondria in the mid-piece via the oxidative phosphorylation system¹⁵. As a result, because in some species sperm velocity predicts fertilization success^{16,17}, it is often assumed that sperm morphology influences sperm swimming velocity and thus fertilization success¹⁸. However, as far as we are aware there is no direct evidence for a relationship between sperm phenotype and sperm swimming velocity, either within or between species.

To determine whether the considerable inter-male variation we observed in zebra finch sperm had any effect on sperm motility, we measured the average path velocity (in $\mu\text{m s}^{-1}$) of sperm for a subset of 105 random males using a Hobson Sperm Tracker¹⁹ but found no relationships between velocity and any sperm morphology trait either individually or in combination (Fig. 2). In the present study it

was not possible to check the repeatability of sperm velocity for the same male; however, in an earlier investigation we found no significant within-male repeatability of sperm velocity across successive ejaculates for the zebra finch¹⁴. This confirms the findings of the present study (using a much larger sample of males) that despite a very high degree of within-male repeatability of flagellum and mid-piece size, there was no association between sperm design and sperm velocity. Sperm motility is complex, and although *in vitro* assays of sperm performance can in some cases predict fertilization success^{16,17}, in general the motility and velocity of sperm will depend very much on the environmental conditions they experience, which for internally fertilizing species is the female reproductive tract⁸.

The degree of inter-male variation in zebra finch sperm design that we observed was extremely high (Fig. 1), and the pattern of variation was surprising. In terms of sperm design we expected a positive, isometric or allometric relationship between flagellum and mid-piece length. Indeed, a positive relationship between these variables exists across passerine species (Fig. 3a), but within the zebra finch the phenotypic correlation between these traits was negative ($r = -0.165$, d.f. = 911, $P < 0.05$), although visual inspection of the data suggests the existence of a positive allometric relationship, albeit with considerable noise (Fig. 3b). Comparison with a small number of samples from wild zebra finches suggests that the observed pattern of variation is not an artefact of captivity or domestication (Fig. 3b).

To elucidate the basis for this inter-male sperm variation we examined the quantitative genetics of flagellum and mid-piece length. Some earlier studies of sperm traits in mice provided inflated heritability estimates³ because they were unable to separate direct genetic and 'maternal effects' in their analyses. However, the model that we used here distinguishes between direct and maternal effects and therefore avoids this problem (see Methods). It is important to clarify what is meant by maternal effects because the term has been

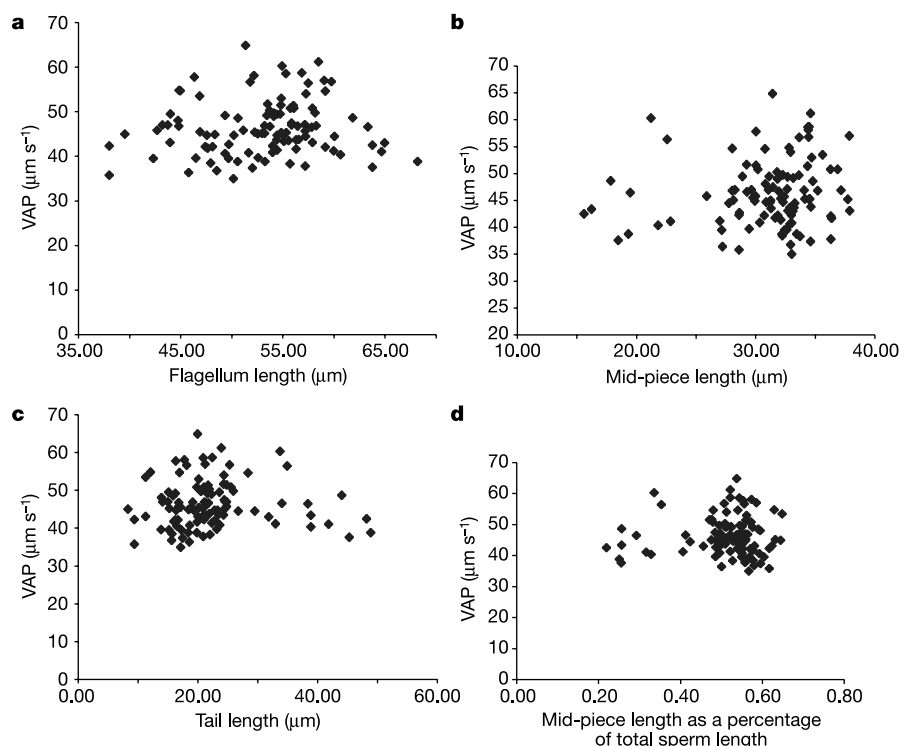


Figure 2 Relationships between zebra finch sperm velocity (VAP) and sperm traits. **a–d**, Relationship between VAP and flagellum length ($r = 0.094$, $P = 0.341$) (**a**), mid-piece length ($r = 0.139$, $P = 0.157$) (**b**), tail length (the portion of the flagellum beyond the mid-piece; see Fig. 1) ($r = -0.007$, $P = 0.945$) (**c**) and mid-piece as a percentage

of flagellum length ($r = 0.032$, $P = 0.746$) (**d**). Degrees of freedom = 103 in all cases. There were no significant correlations between combinations of traits using multiple regression analyses.

Table 1 Data for zebra finch sperm traits

Trait	Mean ($\pm$ s.d.)	CVP	CVA (total)	CVA (direct)	CVA (indirect)	CVR	Heritability (direct genetic effect)	Indirect genetic effect
Head length (μ m)	11.02 $\pm$ 0.61	5.56	3.84	3.84	–	4.02	0.48	–
Mid-piece (μ m)	32.94 $\pm$ 4.36	13.23	9.71	8.84	4.03	8.74	0.45	0.09
Flagellum (μ m)	53.59 $\pm$ 5.11	9.53	8.24	7.49	3.44	4.52	0.62	0.13

CVP, phenotypic coefficient of variation; CVA (total), total additive genetic coefficient of variation; CVA (direct), direct additive genetic coefficient of variation; CVA (indirect), maternal additive genetic coefficient of variation (note that no indirect genetic effect on sperm head length was detected); CVR, environment coefficient of variation. Estimates are based on MTDREML variance component estimates and using an animal model.

Table 2 Genetic correlation matrix for direct and indirect effects on zebra finch sperm traits

	Head	Flagellum (direct)	Flagellum (indirect)	Mid-piece (direct)	Mid-piece (indirect)
Head	0.485 (0.03)	0.127 (0.08)	0.447 (0.23)	–0.185 (0.12)	0.054 (0.05)
Flagellum (direct)	–	0.599 (0.12)	0.258 (0.15)	– 0.505 (0.16)	– 0.284 (0.12)
Flagellum (indirect)	–	–	0.098 (0.05)	0.036 (0.04)	– 0.965 (0.11)
Mid-piece (direct)	–	–	–	0.463 (0.09)	0.195 (0.12)
Mid-piece (indirect)	–	–	–	–	0.092 (0.04)

Diagonal values are heritabilities; non-diagonal values are genetic correlations. Values in parentheses are standard errors; values in bold are statistically significant ($P < 0.05$).

used in several different ways. Maternal effects can be entirely environmental, entirely genetic, or a combination of both environmental and genetic factors. To avoid ambiguity we follow ref. 20 and refer to entirely genetic maternal effects as ‘indirect genetic effects’ and additive genetic effects as ‘direct genetic effects’. Our model (see Methods) estimates only the direct genetic effects and the indirect genetic effects, and specifically excludes maternal environmental effects.

All sperm components were heritable (Table 1) and apart from sperm head length also showed significant indirect genetic effects

(Table 2). Indirect genetic effects on sperm have been suggested by other studies, especially for the mitochondria in the mid-piece, as these are inherited maternally^{21,22}. However, the existence of significant indirect genetic effects for the flagellum (Table 2) was unexpected. If sperm development is influenced by both nuclear and mitochondrial genomes the inevitable antagonistic coevolution may constrain optimal sperm design. Because males are unable to pass maternally inherited traits on to their sons, traits with a maternally inherited component can respond to selection only through the female line and in many instances may be an evolutionary dead end^{20–23}.

Although our analysis suggests that the indirect effects we detected are entirely genetic, quantitative genetics techniques do not allow us to completely rule out the existence of non-genetic maternal effects, mediated, for example, by differential investment in eggs. Because differential endocrine and antioxidant investment in eggs with laying order has been found in the zebra finch²⁴, we tested whether laying order had any effect on sperm flagellum or mid-piece length, but found no such effects despite our extensive data set ($n > 900$ clutches). Although this suggests that little if any of the variance in flagellum or mid-piece length is mediated via non-genetic maternal effects, we cannot completely exclude other possible egg effects.

The genetic correlations between flagellum length and mid-piece length in the zebra finch were negative and pronounced (Table 2), indicating that the alleles for these traits are either antagonistically pleiotropic or in linkage disequilibrium. Another possibility is that if maternal environmental effects on sperm phenotype are small, as we suggest, sperm construction may involve an interaction between the mitochondrial and nuclear genomes. Negative genetic correlations between direct and indirect effects are common, variable across traits and species, and unexplained²⁵, but they are biologically important in the present context because like indirect effects, they can constrain the evolution of optimal sperm design²⁰. For example, the negative genetic correlation between flagellum (direct) and mid-piece (indirect) length (Table 2) indicates that if a male obtains genes from both parents that make his sperm flagellum longer, he will also get genes from his mother that make his sperm mid-piece shorter (and vice versa). Similarly, the pronounced negative genetic correlation between flagellum (indirect) and mid-piece (indirect) length (Table 2) indicates that genes a male obtains from his mother that make his sperm flagellum longer will make his sperm mid-piece shorter (and vice versa). Indeed, this particular negative genetic correlation is so close to 1 it suggests a powerful genetic constraint on sperm design. However, the way that these kinds of genetic processes influence the developmental pathways of sperm is not clear.

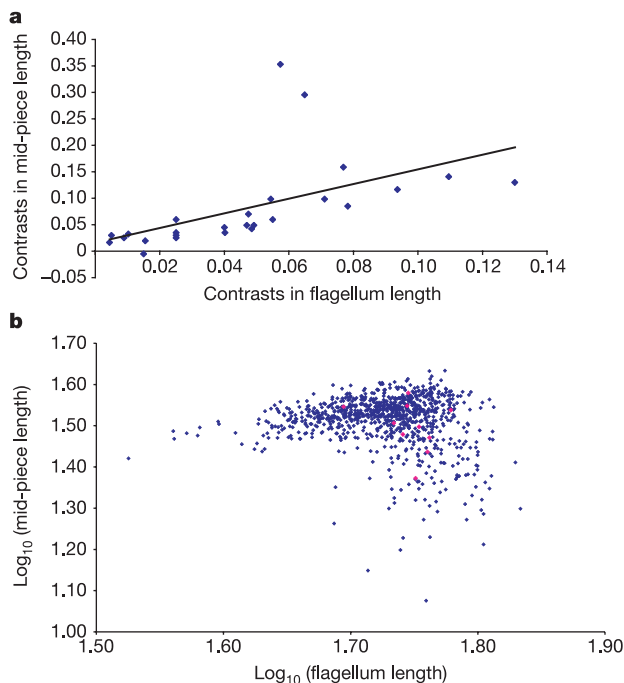


Figure 3 Relationship between sperm flagellum and mid-piece length. **a**, Relationship between contrasts in flagellum and mid-piece length for 29 passerine species controlling for phylogeny using CAIC³⁰: the relationship is significant ($F_{1,26} = 43.95$, $P < 0.0001$; regression equation: $\log(\text{mid-piece}) = 1.61 \times \log(\text{flagellum})$; ($r^2 = 0.64$)). **b**, Log–log plot of sperm flagellum length versus mid-piece length for 913 zebra finches. The phenotypic correlation is negative ($r = -0.165$, 911 d.f., $P < 0.05$) and the slope is significantly less than 1: 0.876 ± 0.0015 (standard error) ($t = 592.0$, $P < 0.0001$). Coloured points ($n = 10$) are data from wild zebra finches and are not included in the analysis.

A possible explanation for our findings is as follows. The marked inter-male variation in sperm design in the zebra finch indicates that there has been little or insufficient selection to break the various negative genetic correlations between sperm traits. Post-copulatory sexual selection is likely to be an important selective force influencing sperm design^{1,2}, but because sperm competition in the zebra finch is not particularly intense⁶ this may account for the observed variation. Sperm design in passerine birds (see Fig. 1) seems to be highly conserved²⁶, and it will be interesting to establish whether in more promiscuous species the negative genetic correlations between flagellum and mid-piece length that we observed are conserved or broken²⁷, and whether the phenotypic variation in sperm design is less than in the zebra finch. □

Methods

Study animals

We used domesticated, wild-type zebra finches (part of an outbred study population based at the University of Sheffield since 1985). Birds were maintained on a 14/10 h light/dark regime at ~20 °C on a standard *ad libitum* zebra finch seed diet and water, which was supplemented with hard-boiled egg and soaked seed during breeding. We used a combined full-sibling and half-sibling breeding design^{12,13}; 81 sires produced two sons from each of six dams, a total of 972 male offspring (entire pedigree comprising 1,526 individuals). Sires were unrelated to each other as far back as grandparents, dams were unrelated to their male partner (back as far as grandparents) and all the females paired to a particular breeding male were unrelated to each other (as far back as grandparents). However, sisters were used (no more than two full sisters) but paired to different males. Each female was paired twice, once each with a different male to increase our ability to detect maternal effects. Pairs in individual cages (50 × 45 × 46 cm high) were established in batches of ~30 over 35 months, producing 15 cohorts each of 100–120 individuals. Eggs were numbered and randomly assigned to foster parents on an individual basis to help separate maternal genetic from maternal environmental effects. Chicks were individually marked soon after hatching and removed from foster parents aged 37 days; their sex (from plumage characteristics) and body mass (referred to as fledging mass) were recorded (between 14:00–16:00 local time). The study focused on male offspring. At sexual maturity (100 d (ref. 28)) various traits were measured relative to a period of exercise in which birds flew ~4 km day⁻¹ for 42 days aged 168–203 days to minimize any effects of a lack of exercise or overeating. After the period of exercise, sperm were obtained from the ejaculatory duct region of the seminal glomerus by dissection, fixed in 5% formalin and viewed using a Spot Diagnostics (version 3) image analysis system at a magnification of × 400. We obtained data for five sperm from each of 913 out of 972 males. Because the degree of coiling of the mid-piece varied between males we estimated the straight mid-piece length (T), thus $T = (L/d)l$, where $l = \sqrt{d^2 + (2\pi r)^2}$, $d = L/N$, L = the length of the mid-piece and N = the number of complete helix gyres, and where r = the radius from the centre of the sperm flagellum to the centre of the mid-piece helix, measured from electron micrographs and assumed to be the same for all males (3 µm).

Genetic analyses

The model used to estimate (co)variance components was a bivariate linear mixed maternal animal model, considering two traits affected by both direct and maternal genetic effects. In matrix notation:

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} + \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + \begin{bmatrix} M_1 & 0 \\ 0 & M_2 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

where Y_i is the vector of observations for the i th trait; B_i is the unknown vector of fixed effects for the i th trait; u_i is the unknown vector of random animal effects for the i th trait; w_i is the unknown vector of random maternal genetic effects for the i th trait; e_i is the vector of residual environmental random effects for the i th trait; and X_i , Z_i and M_i are known incidence matrices relating observations of the i th trait to fixed effects (cohort, clutch size and dam age, when declared significant, or the general mean otherwise), animal genetic effects and maternal genetic effects, respectively.

It is assumed that

$$\text{var} \begin{bmatrix} u_1 \\ u_2 \\ w_1 \\ w_2 \\ e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} g_{11} & g_{12} & g_{13} & g_{14} & 0 & 0 \\ g_{21} & g_{22} & g_{23} & g_{24} & 0 & 0 \\ g_{31} & g_{32} & g_{33} & g_{34} & 0 & 0 \\ g_{41} & g_{42} & g_{43} & g_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & r_{11} & r_{12} \\ 0 & 0 & 0 & 0 & r_{21} & r_{22} \end{bmatrix}$$

where g_{ij} are elements of G and represent the additive genetic covariance between variables i and j , where $i = 1, 2$ refers to animal direct effects for traits 1 and 2 and $i = 3, 4$ refers to maternal genetic effects for traits 1, 2. Similarly r_{ij} are elements of R , the variance and covariance matrix for residual effects.

Genetic parameters were obtained through an iterative process that estimates G , the genetic (co)variance matrix, which is based on calculating the variance of the animal genetic effects, the variance of the maternal genetic effects and the additive relationship matrix among all the animals in the pedigree file (matrix A). Calculations using this model

are based on the additive genetic effects. No other genetic effects (dominance, epistasis) or permanent environmental effects are considered in this model. In separate analyses we found no significant paternal environmental effects.

Subsets of two traits at a time were analysed. Traits studied were mid-piece length and flagellum length. There were 972 birds present in the data but the number of animal effects (the size of A , the numerator relationship matrix) included in this pedigree file was 1,526. We obtained sperm samples from 913 of the 972 male offspring. The animal model was used for estimating genetic and environmental variance components to estimate heritabilities and genetic correlations. The strategy involved obtaining univariate estimates for genetic and environmental (co)variances using the multiple-trait, derivative-free REML algorithm implemented in MTDREML software and using them as starting values in the bivariate analysis¹³. Convergence criteria were attained when the variance of the simplex values was $\leq 10^{-12}$. It was assumed that the global maximum was obtained when two restarts, using previously converged values as starting values, produced convergence with no changes in the first three decimal places of the F -value¹³. Linear models were used to identify important sources of fixed effect variation for each trait. Terms fitted as fixed effects were dam age (quantitative continuous), clutch size (categories 2–11), number of fertile eggs (categories 1–10) and cohort (categories 1–15). Linear associations among all traits were estimated. No linear associations between sperm traits and environmental traits were found. As a first approach single-trait animal models (using MTDREML) that included identified significant fixed effects, animal effect and residual were used to estimate heritabilities for each trait. We used likelihood ratios to test whether the estimates of heritability and genetic correlation were significantly different from zero²⁹. For single-trait animal models, a test statistic for the additive genetic variance was obtained by computing twice the difference in log likelihoods between a complete model where the additive genetic variance is freely estimated and a reduced model where it is constrained to be zero: likelihood ratio statistic = $2(\log(lc) - \log(lr))$, where lc is the likelihood for the complete model and lr is the likelihood for the reduced model. This likelihood ratio statistic is compared with a χ^2 distribution with one degree of freedom using a one-tailed region of rejection. As a second approach, single-trait animal models that additionally included the maternal effect were used to estimate direct (narrow sense) heritability and maternal heritabilities (additive maternal genetic/total phenotypic variance) and the genetic correlation between these two genetic components for each sperm trait. As for multiple-trait animal models, a test statistic for the genetic covariance was obtained by computing twice the difference in log likelihoods between a complete model where genetic correlation is freely estimated and a reduced model where the genetic covariance is set to zero. This test statistic is compared with a χ^2 distribution with one degree of freedom. The coefficients of variation for the direct and maternal additive genetic effects, for the environmental effect as well as for the phenotypic values were calculated as the square root of the respective variance component divided by the mean for the trait and multiplied by 100.

Received 28 October 2004; accepted 19 January 2005; doi:10.1038/nature03374.

- Parker, G. A. in *Sperm Competition and Sexual Selection* (eds Birkhead, T. R. & Møller, A. P.) 3–54 (Academic, London, 1998).
- Birkhead, T. R. & Pizzari, T. Post-copulatory sexual selection. *Nature Rev. Genet.* **3**, 262–273 (2002).
- Beatty, R. A. The genetics of the mammalian gamete. *Biol. Rev.* **45**, 73–119 (1980).
- Morrow, E. H. & Gage, M. J. G. Artificial selection and heritability of sperm length in *Gryllus bimaculatus*. *Heredity* **87**, 356–362 (2001).
- Simmons, L. W. & Kotiaho, J. S. Evolution of ejaculates: patterns of phenotypic and genotypic variation and condition dependence in sperm competition traits. *Evolution* **56**, 1622–1631 (2002).
- Birkhead, T. R., Burke, T., Zann, R., Hunter, F. M. & Krupa, A. P. Extra-pair paternity and intraspecific brood parasitism in wild zebra finches *Taeniopygia guttata*, revealed by DNA fingerprinting. *Behav. Ecol. Sociobiol.* **27**, 315–324 (1990).
- Cohen, J. & McNaughton, D. C. Spermatozoa: the probable selection of a small population by the genital tract of the female rabbit. *J. Reprod. Fertil.* **39**, 297–310 (1974).
- Birkhead, T. R., Møller, A. P. & Sutherland, W. J. Why do females make it so difficult for males to fertilize their eggs? *J. Theor. Biol.* **161**, 51–60 (1993).
- Jamieson, B. G. M. in *The Male Gamete: From Basic Science to Clinical Applications* (ed. Gagnon, C.) 304–331 (Cache River Press, Vienna, Illinois, 1999).
- Cohen, J. *Reproduction* (Butterworths, London and Boston, 1977).
- Morrow, E. H. & Gage, M. J. G. Consistent significant variation between individual males in spermatozoal morphometry. *J. Zool. (Lond.)* **253**, 147–153 (2001).
- Kruuk, L. E. B. Estimating genetic parameters in natural populations using the ‘animal model’. *Phil. Trans. R. Soc. Lond. B* **359**, 873–890 (2004).
- Boldman, K. G., Kriese, L. A., Van Vleck, L. D., Van Tassell, C. P. & Kachman, S. D. *A Manual for Use of MTDREML. A Set of Programs to Obtain Estimates of Variances and Covariances*. Revised United States Department of Agriculture–Agricultural Research Station (US Meat Animal Research Center, Clay Center, Nebraska, 1995).
- Birkhead, T. R. & Fletcher, F. Male phenotype and ejaculate quality in the zebra finch *Taeniopygia guttata*. *Proc. R. Soc. Lond. B* **262**, 329–344 (1995).
- Cardullo, R. A. & Baltz, J. M. Metabolic regulation in mammalian sperm: mitochondrial volume determines sperm length and flagellar beat frequency. *Cell Motil. Cytoskel.* **19**, 180–188 (1991).
- Froman, D. P. & Feltmann, A. J. Sperm mobility: A quantitative trait in the domestic fowl (*Gallus domesticus*). *Biol. Reprod.* **58**, 379–384 (1998).
- Birkhead, T. R., Martinez, J. G., Burke, T. & Froman, D. P. Sperm mobility determines the outcome of sperm competition in the domestic fowl. *Proc. R. Soc. Lond. B* **266**, 1759–1764 (1999).
- Anderson, M. J. & Dixon, A. F. Sperm competition: Motility and the midpiece in primates. *Nature* **416**, 496 (2002).
- Birkhead, T. R., Fletcher, F., Pellatt, E. J. & Staples, A. Ejaculate quality and the success of extra-pair copulations in the zebra finch. *Nature* **377**, 422–423 (1995).
- Wolf, J. B., Brodie, E. D., Cheverud, J. M., Moore, A. J. & Wade, M. J. Evolutionary consequences of indirect genetic effects. *Trends Ecol. Evol.* **13**, 64–69 (1998).
- Frank, S. A. & Hurst, L. D. Mitochondria and male disease. *Nature* **383**, 224 (1996).

22. Gemmel, N. J., Metcalf, V. J. & Allendorf, F. W. Mother's curse: the effect of mtDNA on individual fitness and population viability. *Trends Ecol. Evol.* **19**, 238–244 (2004).
23. Zeh, J. Sexy sons: a dead end of cytoplasmic genes. *Proc. R. Soc. Lond. B (Suppl.)* **271**, S306–S309 (2004).
24. Royle, N. J., Surai, P. F. & Hartley, I. R. The effect of variation in dietary intake on maternal deposition of antioxidants in zebra finch eggs. *Funct. Ecol.* **17**, 472–481 (2003).
25. Cheverud, J. M., Leamy, L. J., Atchley, W. R. & Rutledge, J. J. Quantitative genetics and the evolution of ontogeny I. Ontogenetic changes in quantitative genetic variance components in randombred mice. *Genet. Res.* **42**, 65–75 (1983).
26. McFarlane, R. W. The taxonomic significance of avian sperm. *Proc. 13th Int. Orn. Congr.* **1**, 91–102 (1963).
27. Roff, D. A. The evolution of the G-matrix: selection or drift? *Heredity* **84**, 135–142 (2000).
28. Zann, R. A. *The Zebra Finch: A Synthesis of Field and Laboratory Studies* (Oxford Univ. Press, Oxford, 1996).
29. Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits* (Sinauer Associates, Sunderland, Massachusetts, 1998).
30. Purvis, A. & Rambaut, A. Comparative analysis by independent contrasts (CAIC): an Apple Macintosh application for analysing comparative data. *Comput. Appl. Biosci.* **11**, 247–251 (1995).

Acknowledgements We are grateful to the following for technical assistance: A. Bamford, H. Basford, S. Bawden, S. Bradish, L. Birkhead, A. Blake, M. Hudson, K. Hutchence, R. Linacre, B. Mappin, A. MacDonald, the late O. Scott-Roberts, D. Rose, J. Shutt, K. Swinglehurst, L. Turton and E. Varsey. We thank F. M. Hunter, I. M. Matthews, N. Roddis and P. Young for help with the project, and A. Beckerman, J. D. Biggins, D. Coltman, J. Cummins, C. Haley, L. Keller, A. Moore, B. Sheldon, J. Slate, J. St John and D. Woolley for advice and/or comments. This study was funded by a NERC research grant to T.R.B.

Authors' contributions T.R.B. managed the entire project and wrote the manuscript; E.J.P. managed the birds; P.B. and R.Y. measured the sperm; and H.C.-J. conducted the genetic analyses.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.R.B. (T.R.Birkhead@sheffield.ac.uk).

Optimal eye movement strategies in visual search

Jiri Najemnik & Wilson S. Geisler

Center for Perceptual Systems and Department of Psychology, University of Texas at Austin, Austin, Texas 78712, USA

To perform visual search, humans, like many mammals, encode a large field of view with retinas having variable spatial resolution, and then use high-speed eye movements to direct the highest-resolution region, the fovea, towards potential target locations^{1,2}. Good search performance is essential for survival, and hence mammals may have evolved efficient strategies for selecting fixation locations. Here we address two questions: what are the optimal eye movement strategies for a foveated visual system faced with the problem of finding a target in a cluttered environment, and do humans employ optimal eye movement strategies during a search? We derive the ideal bayesian observer^{3–6} for search tasks in which a target is embedded at an unknown location within a random background that has the spectral characteristics of natural scenes⁷. Our ideal searcher uses precise knowledge about the statistics of the scenes in which the target is embedded, and about its own visual system, to make eye movements that gain the most information about target location. We find that humans achieve nearly optimal search performance, even though humans integrate information poorly across fixations^{8–10}. Analysis of the ideal searcher reveals that there is little benefit from perfect integration across fixations—much more important is efficient processing of information on each fixation. Apparently, evolution has exploited this fact to achieve efficient eye movement strategies with minimal neural resources devoted to memory.

Recent decades have seen considerable progress in understanding visual search^{11,12}, eye movements^{1,2,13} and active robotic vision¹⁴; however, there is no formal theory of optimal eye movement strategies in conducting visual search. Such a theory would provide insight into the design requirements for effective control of eye movements and attention, and hence could serve as a powerful framework for analysing the behaviour and neurophysiology of eye movements and attention, and for developing robotic applications.

We consider the task of finding a known target that is embedded (added) at a random location in backgrounds of spatial $1/f$ noise, which have the same spatial power spectra as images of natural scenes⁷. Figure 1a shows the target (a spatial sine wave) and a sample of $1/f$ noise.

Not surprisingly, the optimal eye movement strategy depends critically on how the visibility (detectability) of the target varies across the retina. Thus, to specify the ideal searcher, it is necessary first to characterize the visibility maps of the visual system under consideration, for the targets and backgrounds of interest.

To characterize the visibility map for each of the conditions in the search experiment described below, detection accuracy was measured for the sine-wave target as a function of target contrast and background noise contrast, at the 25 spatial locations indicated by the small circles in Fig. 1a. The observer fixated on the centre of the display, which was monitored with an eye tracker, and for each block of trials the target was presented at only a single known location. Figure 1b shows the measurements of detection accuracy (psychometric functions) in the fovea, as a function of target contrast. Each curve is for a different root-mean-squared (r.m.s.) contrast of the noise background, where the r.m.s. contrast is defined as the standard deviation of the pixel luminance divided by the mean luminance. Each of these psychometric functions can

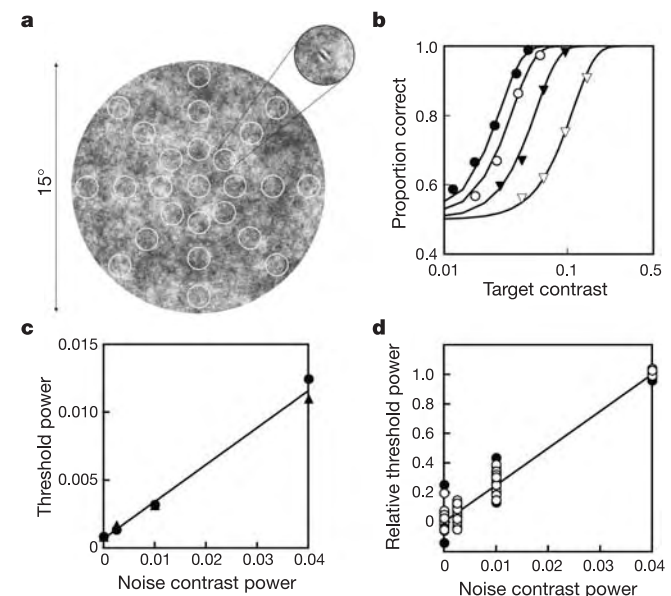


Figure 1 Measurement of the visibility maps. **a**, The target was a windowed sine-wave grating (see inset). Target visibility was measured at each of the locations indicated by the circles (which were not in the stimulus), in a two-interval forced-choice detection task. Each interval was 250 ms, the approximate duration of single fixations in our visual search experiments (280 ms). **b**, Proportion of correct responses in the fovea, as a function of target contrast, for four levels of background noise contrast: filled circles, 0; open circles, 0.05; filled triangles, 0.10; open triangles, 0.20. **c**, Detection threshold power in the fovea for two observers, J.N. (circles) and W.G. (triangles). **d**, Relative threshold power for all measured locations in **a**. These plots were obtained by normalizing the data so that the best-fitting line in each condition had an intercept of 0 and a value of 1.0 at a contrast power of 0.04. Observers: filled circles, J.N.; open circles, W.G.

be summarized by a contrast threshold value, the target contrast that is detected with 82% accuracy, and by a parameter that describes the steepness of the function (see Methods). Figure 1c plots the foveal contrast thresholds (in units of contrast power—the square of the contrast) for two observers. The thresholds fall approximately along a straight line, in agreement with previous studies using white-noise backgrounds^{15–17}. Figure 1d plots the relative threshold (see the figure legend) as a function of noise contrast power for all the retinal locations. The relative thresholds cluster around a straight line, showing that an approximately linear relationship holds for all retinal locations tested; however, there are systematic changes in the slopes and intercepts of the lines with the distance of the target from the fovea (the retinal eccentricity), as shown in Fig. 2a, b. These slope and intercept functions (together with the steepness parameter of the psychometric function) can be used to determine the visibility map for any combination of target and background contrast. Figure 2c shows a cross-section of the maps for two of the conditions in the search experiment described later. Each map specifies, for every retinal eccentricity, the value of a signal-to-noise ratio, d' , which is monotonically related to detection accuracy (see Methods)³.

Now consider the ideal searcher for the relatively simple search task in which the target location is unknown but the stimulus is presented briefly so that no eye movements are possible. In this case, the optimal method is template matching (that is, cross-correlation)^{3,4}. At each potential target location, the retinal image is multiplied by a template of the target and the product is integrated to obtain a template response. If all target locations have equal prior probability, and if the visibility map is flat, then the location with the largest template response is the most likely location of the target (that is, the location with the greatest posterior probability). To compute posterior probabilities when the visibility map is not flat,

template responses are weighted by the visibility at each potential target location.

The temporally extended search task with eye movements introduces two new requirements for the ideal searcher: optimal integration of responses across fixations, and optimal selection of successive fixation locations. The flow diagram of processing performed by the ideal searcher is shown in Fig. 3a. During the first fixation the searcher captures responses from all potential target locations. It then computes the posterior probability that the target is located at each of these locations. If the maximum of the posterior probabilities exceeds a criterion (the criterion determines the error rate), the search stops and the location with the largest posterior probability is reported. If the criterion is not exceeded, the ideal searcher determines the fixation location that will maximize the probability of finding the target after the eye movement is made. It then moves its eyes to that location, and the process repeats.

To integrate responses across fixations optimally, the ideal searcher cumulates the weighted responses from each potential target location:

$$p_i(T) = \frac{\text{prior}(i) \exp\left(\sum_{t=1}^T d'_{ik(t)} W_{ik(t)}\right)}{\sum_{j=1}^n \text{prior}(j) \exp\left(\sum_{t=1}^T d'_{jk(t)} W_{jk(t)}\right)} \quad (1)$$

where t is fixation number, and $d'_{ik(t)}$ and $W_{ik(t)}$ are the visibility and response at display location i when the fixation is at display location $k(t)$. Equation (1) is for the case in which both stimulus noise and internal noise are statistically independent in time (dynamic). Equations for the more complicated case, in which the noise is a mixture of static stimulus noise and dynamic internal noise (like the present experiments), are given in the Supplementary Information. The predictions reported here are for the static case, although predictions for the dynamic case are similar (see Methods).

To compute the optimal next fixation point, $k_{\text{opt}}(T+1)$, the ideal searcher considers each possible next fixation and picks the location that, given its knowledge of the current posterior probabilities and visibility map, will maximize the probability of correctly identifying the location of the target after the fixation:

$$k_{\text{opt}}(T+1) = \arg \max_{k(T+1)} \left(\sum_{i=1}^n p_i(T) p(C|i, k(T+1)) \right) \quad (2)$$

Maximizing accuracy is the relevant goal in the present task; other goals (such as minimizing entropy) have been explored for certain computer vision tasks¹⁸. Explicit expressions for equation (2) are derived in the Supplementary Information.

Given the prior probabilities of possible target locations and the visibility maps, which specify all the relevant values of d' , equations (1) and (2) can be used to simulate the behaviour of the ideal searcher (see Methods). Figure 3b shows a sequence of ideal fixations. The ideal searcher performs a rather random-looking search pattern, although it is in fact a highly principled search that reflects the specific properties of the stimulus and the visibility map. This figure also shows how posterior probabilities across the display evolve over time. The location of the target is at the peak in the posterior probability map seen downward of the left of centre.

The ideal searcher shows several other interesting qualitative behaviours. First, it sometimes makes fixations to the display location with the maximum posterior probability of containing the target (MAP fixations, where MAP is short for maximum *a posteriori*), and sometimes to a location near the centroid of a cluster of locations where the posterior probabilities are high ('centre-of-gravity' fixations). Both MAP and centre-of-gravity fixations have been observed in human visual search^{19–21}. Second, the saccade lengths of the ideal searcher tend to be moderate in size, because posterior probabilities at nearby locations are pushed down and posterior probabilities at distant locations tend not to jump up (see Fig. 3b). Human saccade lengths also tend to be moderate in

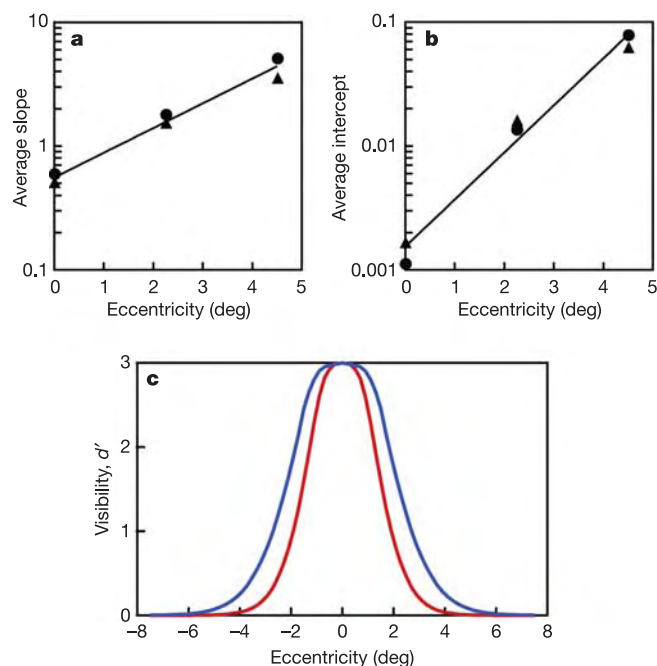


Figure 2 Representation of the visibility maps. **a**, **b**, The average slopes (**a**) and intercepts (**b**) of the threshold functions shown in Fig. 1c, d as a function of retinal eccentricity. Observers: circles, J.N.; triangles, W.G. **c**, The signal-to-noise ratio, d' , for target detection across the visual field (visibility maps) for two of the conditions in the search experiment. The red function is for a background contrast of 0.05 and a target contrast of 0.07. The broader blue function is for a background contrast of 0.20 and a target contrast of 0.19. Note that visibility reaches a peak in the centre of the fovea and decreases rapidly with retinal eccentricity.

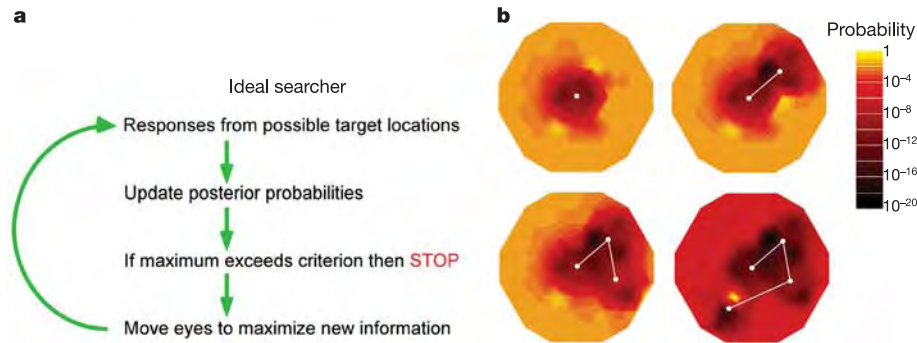


Figure 3 Ideal searcher. **a**, Flow chart for the ideal searcher. **b**, A typical sequence of fixations for the ideal searcher, for which the initial fixation is at the centre of the display. The temperature plots show the posterior probability map after each fixation. There is an

elevated posterior probability at the target location (downward of the left of centre), but initially the eye is drawn to other locations. Posterior probabilities are suppressed in the neighbourhood of each fixation, creating 'inhibition of return'.

size (see later). Third, the ideal searcher tends not to fixate display locations that it has recently fixated ('inhibition of return'), again because nearby posterior probabilities are pushed down (see Fig. 3b). Fourth, the ideal searcher sometimes makes long saccades into regions where the posterior probabilities are low, followed by a return saccade to a region with higher probabilities. It performs these eye movements because excluding an unlikely region that has not yet been inspected is sometimes the best chance for increasing the posterior probabilities in the more likely regions. It is unknown whether humans perform these 'exclusion saccades' in visual search, although a related type of saccade is predicted for optimal eye movements in reading²².

To compare human and ideal search quantitatively, we measured search performance for the sine-wave target randomly embedded at one of 85 locations tiling the 15° diameter display in a triangular array. Measurements were made for two levels of 1/f noise contrast (0.05 and 0.2) and for six levels of target visibility in the fovea ($d' = 3, 3.5, 4, 5, 6$ and 7). The visibilities were set by using the results from the detection experiment (see Figs 1 and 2). The data points in Fig. 4a show the median number of fixations required for two human observers to find the target. (We obtained a similar search performance for a third observer naive to the aims of the study, although we do not have visibility maps for that observer.) As can be seen, search performance improves as the visibility of the target increases and is better in the high-noise condition (presumably because the visibility maps are broader; see Fig. 2c). The solid curves show the predictions of the ideal searcher with the same visibility maps as the human observers. Figure 4b shows how human and ideal search performance varies with location of the target in the display, for all 12 stimulus conditions. The results in Fig. 4a, b imply that humans are remarkably efficient at visual search, at least under these conditions, nearly reaching the performance of the ideal searcher.

The obvious question is: How do humans perform so well? To address this question, consider the three things that the ideal searcher does in an optimal fashion: parallel detection, integration of information across fixations, and selection of the next fixation location. With regard to parallel detection, there is evidence that humans are efficient at finding targets in noise under conditions in which visibility is equal at all potential target locations^{4,23}. Further, in brief presentations that do not allow eye movements, humans often process multiple target locations in parallel with equal efficiency^{11,23,24}. Our results indicate strongly that humans are able to perform this kind of efficient parallel processing in complex extended search tasks.

With regard to the integration of information across fixations, there is evidence that humans are not very efficient⁸⁻¹⁰. So how can they approximate ideal performance? A key insight is provided by

the solid curve in Fig. 4c, which plots the average posterior probability at the target location as a function of the number of fixations before the one at which the ideal searcher found the target. The rapid rise in posterior probability implies that there is little to be gained by integrating detailed display or posterior probability information more than one or two fixations into the past. However,

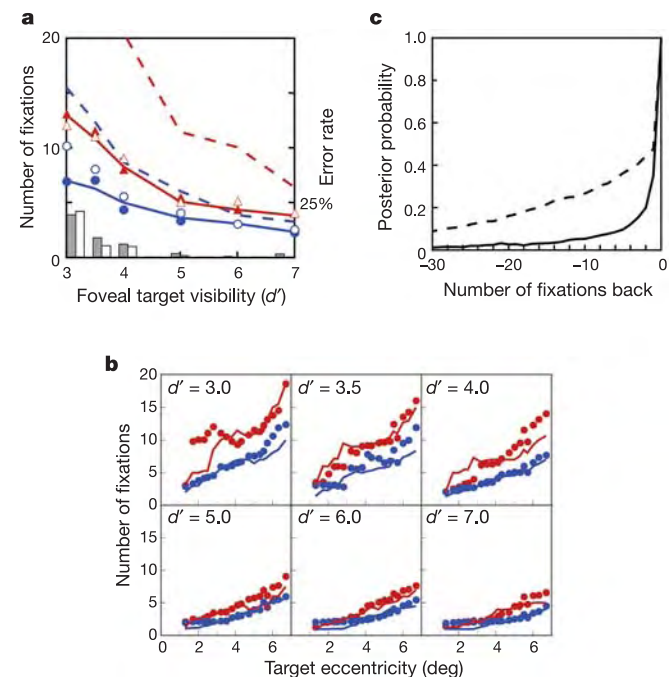


Figure 4 Human versus ideal performance. **a**, Median number of fixations to locate the target correctly, as a function of the target's visibility in the fovea, for two levels of background noise contrast: 0.05 (red lines and triangles) and 0.20 (blue lines and circles). Solid and dashed curves are the predictions of the ideal and random searchers, respectively. The histogram at the bottom shows the error rates of the human observers (grey) and an ideal searcher (white). Observers: filled symbols, J.N.; open symbols, W.G. **b**, Median number of fixations to locate the target correctly as a function of the target's eccentricity and its visibility in the fovea. The symbols are combined results for two observers; the solid curves are the predictions of the ideal searcher. The medians were obtained by binning with a sliding 2° window. The medians are less reliable at small eccentricities because the average number of observations in a bin increases with the square of the eccentricity. Noise contrast: red, 0.05; blue, 0.20. **c**, Posterior probability at the target location as a function of the number of fixations before the one in which the searcher found the target. Solid and dashed curves are the predictions of the ideal and random searchers, respectively.

simulations show that to achieve human performance levels it is necessary to have some coarse memory for past fixation locations so as to reduce the likelihood of returning to the same display region.

With regard to the selection of fixation locations, there is little evidence about human efficiency. To examine this issue we simulated searchers that do not select fixation locations optimally but are otherwise ideal. For example, the dashed curves in Fig. 4a show the performance of a searcher that computes posterior probabilities and integrates information across fixations optimally, but makes random fixations. Humans far outperform this random searcher. They also outperform an enhanced version with the added feature of not fixating the same location twice. The fact that humans outperform these searchers rejects all possible models of visual search where fixation locations are selected at random, with or without replacement. We have also evaluated a sub-optimal searcher that always fixates the location with the maximum posterior probability (the MAP searcher). This searcher performs almost as well as the ideal (0.5–2 fixations slower), and thus MAP fixations alone (no centre-of-gravity fixations) are sufficient to approach the efficiency of human fixation selection.

The fixation selection strategies of the searchers can also be compared by examining eye movement statistics. For example, the distributions of saccade length for the humans and the ideal searcher are similar: both are skewed to the right and reach a peak at about 3°, although the ideal distribution is more skewed than the human. The distribution for the random searcher is symmetrical and reaches a peak at about 7°. The distribution for the MAP searcher is similar to the ideal and the human, but less skewed than the human. We also examined the spatial distribution of fixation locations within the display. For both human and ideal searchers the average distribution of fixation locations across the display has a 'doughnut' shape that reaches a peak at about 5° from the centre. Interestingly, the MAP searcher does not have a doughnut-shaped distribution, indicating that humans are not well modelled as MAP searchers.

The present study is another example of how bayesian ideal observer analysis can provide valuable insight into sensory and perceptual processing^{5,6,22}. The ideal searcher is, in some ways, complementary to existing computational models of visual search^{25,26}. Unlike these previous approaches, it is not a heuristic model that can be applied to arbitrary stimuli but a formal, parameter-free analysis for a particular class of naturalistic stimuli. The ideal searcher is not meant to be a plausible model of human visual search (for example, humans do not have perfect memory), but a rigorous starting point for developing models. It remains to be determined what algorithms the brain uses to achieve near-optimal performance for our stimuli, how those algorithms are implemented in neural circuits, and how well those algorithms perform on a more general range of stimuli (for example, on natural images). Nonetheless, it is clear that humans compute something close to an accurate posterior probability map and then use that map to determine the next fixation location efficiently. They are able to reach near-optimal performance, despite relatively poor memory for visual detail and poor integration of information across fixations, because all that is actually needed is a coarse memory sufficient to support the inhibition of return. □

Methods

Stimuli and apparatus

Stimuli were presented on a calibrated monochrome Image Systems monitor (M21L) with a white phosphor (P-104) and resolution 1024 × 768 at 60 Hz. The target was a 6 cycles deg⁻¹ sine-wave grating, tilted 45° to the left and windowed by a symmetrical raised cosine (half-height width one cycle). The background was filled with 1/f noise at a mean luminance of 40 cd m⁻²; the display outside the background was set to the mean luminance. The 1/f noise was created by filtering gaussian white noise, truncating the filtered noise waveform to ±2 s.d., and scaling to obtain the desired r.m.s. amplitude. Eye position was measured with a Forward Technologies SRI Version 6 dual Purkinje eye tracker. Head position was maintained with a bite bar and headrest.

Detection experiment

On each trial, the observer fixated a dot at the centre of the display (there was no fixation dot for foveal measurements). The fixation dot disappeared at the onset of the test stimuli, which consisted of two 250-ms intervals separated by 500 ms. One interval contained background noise alone, the other a different random sample of background noise with the target added. The observer judged which interval contained the target. If the observer's fixation was not within 0.75° of the fixation dot when the test stimuli appeared, the trial was discarded. Within a session the target always appeared in the same location, which was indicated after each trial. Sixteen blocks (four target contrasts × four noise levels) of 32 trials were run in each session. The data for each noise level, in each session, were fitted with a Weibull function:

$$f(c) = 0.5 + 0.5[1 - \exp\{- (c/c_T)^s\}] \quad (3)$$

Both the 82% correct threshold parameter c_T and the steepness parameter s were estimated by using maximum-likelihood methods.

Search experiment

On each trial, the observer first fixated a dot at the centre of the display and then initiated the trial with a button press. The fixation dot disappeared immediately and a random time later (100–500 ms) the search display appeared. The observer's task was to find the target as rapidly as possible and to press a button as soon as the target was located. The observer then indicated the judged target location by fixating that location and pressing the button again. The response was considered correct if the eye position at the time of the second button press was closer to the target location than to any other potential target location. Each session consisted of 6 blocks of 32 search trials. In each session the background noise contrast was fixed. Before each block, the eye tracker was calibrated and the search target for that block was shown on a uniform background at the centre of the display. The set of 12 conditions was repeated 6 times in a counterbalanced fashion.

Search simulations

The solid curves in Figs 1c, 1d, 2a and 2b describe how the threshold parameter $c_T(e_n, \epsilon)$ varied with noise contrast power e_n and eccentricity e . The steepness parameter of the psychometric functions varied with eccentricity from a value of 2 in the fovea to more than 4 in periphery, and was well approximated by $s(e) = 2.8e/(e + 0.8) + 2$. These descriptive models were substituted into equation (3) to obtain a formula for detection accuracy: $f(c, e_n, \epsilon)$. The visibility map was obtained by taking the inverse standard normal integral of the accuracy: $d'(c, e_n, \epsilon) = \sqrt{2} \Phi^{-1}(f(c, e_n, \epsilon))$. The factor 2 takes into account that there were two intervals in the forced choice detection task, but (effectively) only one interval in each fixation of the search task³. The visibility maps were obtained by averaging across all directions; in fact, the visibility maps are not radially symmetric but fall off rather faster in the vertical direction than in the horizontal direction. Using the radial average has no effect on the predictions described here.

Here we describe the simulations for the case in which the external and internal noise is statistically independent (dynamic) in time; the Supplementary Information describes the case in which the external noise is static. We note first that the template responses at each location can be given an expected value of 0.5 when the target is at that location, and -0.5 otherwise. This has no effect on the predictions as long as we add enough noise to the simulated template responses to make the values of d' exactly those given by the visibility map. The six steps for each simulated search trial were as follows. First, fixation began at the centre of the display (as in the search experiment). Second, a target location was selected at random (prior(i) = 1/85); 0.5 was added to that location and -0.5 to all other locations. Third, a gaussian noise sample was generated for each of the 85 potential target locations. The standard deviation of each noise sample was set to be consistent with the value of the visibility map at that location: $\sigma = 1/d'$. Fourth, the posterior probability for each potential target location was calculated from equation (1). Fifth, if the maximum posterior probability exceeded a criterion, the search stopped. The criterion for each condition was picked so that the error rate of the ideal searcher approximated that of the humans. Sixth, for the ideal searcher, equation (2) was evaluated to select the next fixation location; for the random searcher, the next fixation location was selected at random. The process then jumped back to the third step. Note that the specific characteristics of the target and 1/f noise enter the simulation through the visibility maps. Also note that although 1/f noise is spatially correlated, we verified by simulation that the template responses were effectively uncorrelated. The ideal searcher performs slightly better in the dynamic case than in the static case, because it gains more information on each fixation, but the predicted curves are nearly parallel.

Received 8 December 2004; accepted 27 January 2005; doi:10.1038/nature03390.

1. Carpenter, R. H. S. (ed.) *Eye Movements* (Macmillan, London, 1991).
2. Liversedge, S. P. & Findley, J. M. Saccadic eye movements and cognition. *Trends Cogn. Sci.* **4**, 6–14 (2000).
3. Green, D. M. & Swets, J. A. *Signal Detection Theory and Psychophysics* (Wiley, New York, 1966).
4. Burgess, A. E. & Ghandeharian, H. Visual signal detection. II. Effect of signal-location identification. *J. Opt. Soc. Am. A* **1**, 906–910 (1984).
5. Geisler, W. S. & Diehl, R. L. A Bayesian approach to the evolution of perceptual and cognitive systems. *Cogn. Sci.* **27**, 379–402 (2003).
6. Kersten, D., Mamassian, P. & Yuille, A. L. Object perception as Bayesian inference. *Annu. Rev. Psychol.* **55**, 271–304 (2004).
7. Field, D. J. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A* **4**, 2379–2394 (1987).
8. Irwin, D. E. Information integration across saccadic eye movement. *Cognit. Psychol.* **23**, 420–458 (1991).
9. Hayhoe, M. M., Bensinger, D. G. & Ballard, D. H. Task constraints in visual working memory. *Vision Res.* **38**, 125–137 (1998).

10. Rensink, R. A. Change detection. *Annu. Rev. Psychol.* **53**, 245–277 (2002).
11. Palmer, J., Verghese, P. & Pavel, M. The psychophysics of visual search. *Vision Res.* **40**, 1227–1268 (2000).
12. Wolfe, J. M. in *Attention* (ed. Pashler, H.) 13–74 (Psychology Press, Hove, East Sussex, 1998).
13. Schall, J. D. in *The Visual Neurosciences* (eds Chalupa, L. M. & Werner, J. S.) 1369–1390 (MIT Press, Cambridge, Massachusetts, 2004).
14. Blake, A. & Yuille, A. L. (eds) *Active Vision* (MIT Press, Cambridge, Massachusetts, 1992).
15. Burgess, A. E., Wagner, R. F., Jennings, R. J. & Barlow, H. B. Efficiency of human visual signal discrimination. *Science* **214**, 93–94 (1981).
16. Pelli, D. G. & Farell, B. Why use noise? *J. Opt. Soc. Am. A* **16**, 647–653 (1999).
17. Lu, Z.-L. & Doshier, B. A. Characterizing human perceptual inefficiencies with equivalent internal noise. *J. Opt. Soc. Am. A* **16**, 764–778 (1999).
18. Geman, D. & Jedynak, B. An active testing model for tracking roads in satellite images. *IEEE Trans. Pattern Anal. Mach. Intell.* **18**, 1–14 (1996).
19. Rajashekar, U., Cormack, L. K. & Bovik, A. C. in *Eye Tracking Research & Applications* (ed. Duchowski, A. T.) 119–123 (ACM SIGGRAPH, New Orleans, 2002).
20. Findley, J. M. Global processing for saccadic eye movements. *Vision Res.* **22**, 1033–1045 (1982).
21. Zelinsky, G. J., Rao, R. P., Hayhoe, M. M. & Ballard, D. H. Eye movements reveal the spatio-temporal dynamics of visual search. *Psychol. Sci.* **8**, 448–453 (1997).
22. Legge, G. E., Hooven, T. A., Klitz, T. S., Mansfield, J. G. & Tjan, B. S. Mr. Chips 2002: New insights from an ideal observer model of reading. *Vision Res.* **42**, 2219–2234 (2002).
23. Eckstein, M. P., Butcher, B. R. & Stone, L. S. Quantifying the performance limits of human saccadic targeting during visual search. *Perception* **30**, 1389–1401 (2001).
24. Eckstein, M. P., Thomas, J. P., Palmer, J. & Shimozaki, S. S. A signal detection model predicts the effects of set size on visual search accuracy for feature, conjunction, triple conjunction, and disjunction displays. *Percept. Psychophys.* **62**, 425–451 (2000).
25. Itti, L. & Koch, C. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Res.* **40**, 11–46 (2000).
26. Rao, R. P. N., Zelinsky, G. J., Hayhoe, M. M. & Ballard, D. H. Eye movements in iconic visual search. *Vision Res.* **42**, 1447–1463 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank R.F. Murray for helpful discussions, and J. Perry, L. Stern and C. Creeger for technical assistance. This work was supported by the National Eye Institute, NIH.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to W.S.G. (geisler@psy.utexas.edu).

The first cleavage of the mouse zygote predicts the blastocyst axis

Berenika Plusa^{1,2}, Anna-Katerina Hadjantonakis^{3,*}, Dionne Gray¹, Karolina Piotrowska-Nitsche¹, Agnieszka Jedrusik¹, Virginia E. Papaioannou³, David M. Glover² & Magdalena Zernicka-Goetz^{1,2}

¹Wellcome Trust and Cancer Research UK Gurdon Institute, University of Cambridge, Tennis Court Road, Cambridge CB2 1QR, UK

²Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK

³Department of Genetics and Development, College of Physicians and Surgeons, Columbia University, 701 West 168th Street, New York, New York 10032, USA

* Present address: Developmental Biology Program, Sloan-Kettering Institute, 1275 York Avenue, New York, New York 10021, USA

One of the unanswered questions in mammalian development is how the embryonic–abembryonic axis of the blastocyst is first established. It is possible that the first cleavage division contributes to this process, because in most mouse embryos the progeny of one two-cell blastomere primarily populate the embryonic part of the blastocyst and the progeny of its sister populate the abembryonic part^{1–4}. However, it is not known whether the embryonic–abembryonic axis is set up by the first cleavage itself, by polarity in the oocyte that then sets the first cleavage plane with respect to the animal pole, or indeed whether it can be divorced entirely from the first cleavage and established

in relation to the animal pole. Here we test the importance of the orientation of the first cleavage by imposing an elongated shape on the zygote so that the division no longer passes close to the animal pole, marked by the second polar body. Non-invasive lineage tracing shows that even when the first cleavage occurs along the short axis imposed by this experimental treatment, the progeny of the resulting two-cell blastomeres tend to populate the respective embryonic and abembryonic parts of the blastocyst. Thus, the first cleavage contributes to breaking the symmetry of the embryo, generating blastomeres with different developmental characteristics.

It has been shown, either by marking cells directly or by relating their position to external markers, that the majority of first cleavage divisions generate one blastomere whose progeny will primarily populate the embryonic part of the blastocyst (polar trophectoderm and deeper cells of the inner cell mass) and another whose progeny populate the abembryonic part (mural trophectoderm and the more superficial inner cell mass)^{1–4}. More recently, further lineage tracing experiments have clarified this relationship and shown that, by the four-cell stage of normally developing embryos, blastomeres not only have specific fates but can also differ in their developmental potential^{5,6}. Static observations and a series of micromanipulation experiments to relocate the animal pole of the mouse zygote indicate that in normal circumstances the first cleavage usually passes close to this pole^{7,8}. A second marker of the first cleavage, the sperm entry site¹, reflects positioning of the spindle in relation to a slight flattening of the zygote at fertilization⁹. Together these observations raise the question of whether it is the spatial organization of the zygote with respect to the animal pole or, because cleavage orientation can reflect the random site of sperm entry, cleavage itself that influences the subsequent development of the embryonic–abembryonic axis.

To approach this question, we first developed a better method for accurate three-dimensional examination of the first cleavage orientation in a random population of zygotes. We achieved this by time-lapse observations of the first cleavage on multiple focal planes at each time point. In addition, to facilitate analysis of the position of the first cleavage plane in respect to pronuclei (and the polar body (PB)), we followed this cleavage in a newly developed transgenic line (CAG:H2B–EGFP) in which chromatin was labelled green by the expression of green fluorescent protein (GFP)-tagged histone H2B (ref. 10; Fig. 1a; Supplementary Movie 1). We found that the female pronucleus was closer to the PB than the male pronucleus in 87.5% ($n = 32$) of zygotes (categories I–III; Fig. 1b) and that in most zygotes (category I, 62.5%) the two pronuclei were aligned with the PB and the cleavage furrow formed within 30° of this plane. In 12.5% of zygotes (category II), the pronuclei were displaced from the PB by between 30° and 90° and the cleavage plane was displaced to a similar extent. In the two remaining categories, cleavage passed within 30–90° of PBs aligned with the pronuclei (category III), and within 30° of PBs not aligned with the pronuclei (category IV). We observed that this most frequent alignment of the differentially condensed¹¹ female and male pronuclei with the PB was maintained into prometaphase. Immunostaining of fixed zygotes to reveal CpG-methylated female chromatin showed that, in agreement with previous studies^{11,12} and contrary to a recent proposition¹³, the parental sets of chromosomes did not undergo mixing and adopted an arrangement consistent with our time-lapse studies that was retained into the daughter blastomeres (Fig. 1c).

To characterize the orientation of the cleavage plane further, we performed time-lapse differential interference contrast (DIC) microscopy on zygotes of a different, wild-type mouse strain (Fig. 1d). By observing multiple focal planes we ensured that we could follow cleavage in relation to the location of markers in three dimensions in 96% of all randomly positioned zygotes. We found that the onset of cleavage furrow formation was within 30° of the PB in 70% of zygotes, 14-fold greater than at a lateral position between

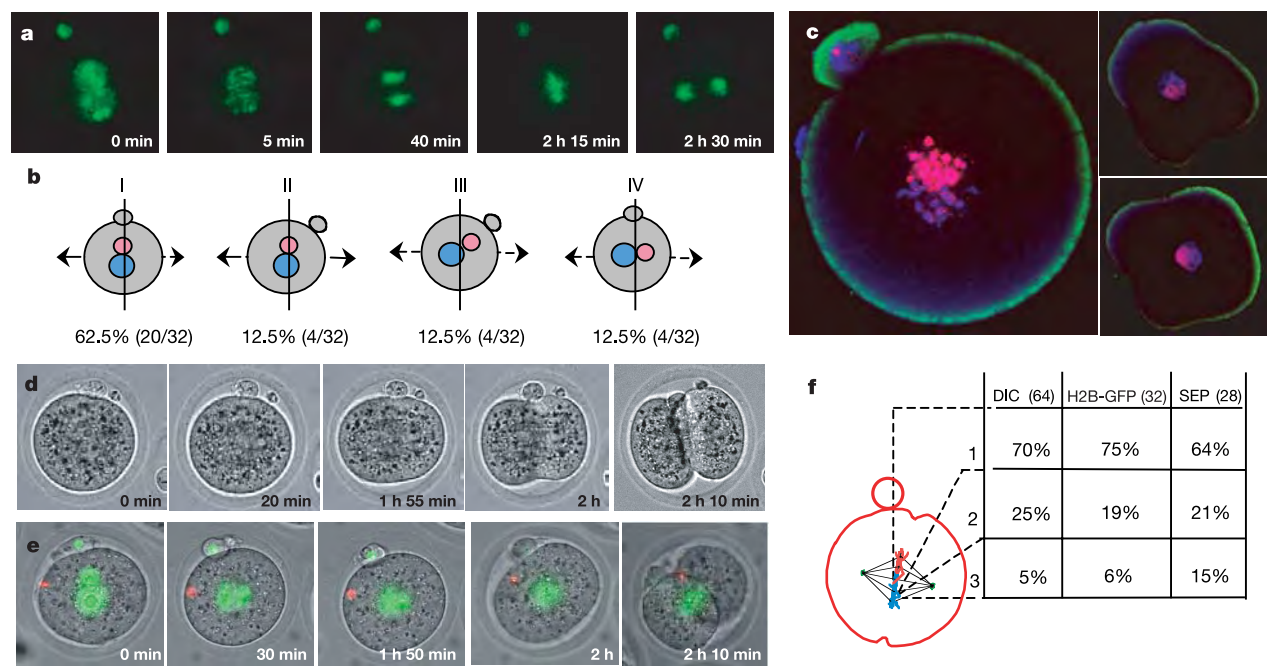


Figure 1 Orientation of the first cleavage division. **a**, One zygote expressing H2B–EGFP from 32 embryos observed by time-lapse imaging, to note the positions of maternal and paternal chromosomes before chromosome condensation, during prometaphase, anaphase and cytokinesis. **b**, Four cleavage categories of zygotes from **a** (vertical line, cleavage plane; horizontal arrows, direction of cell separation): I, male and female pronuclei aligned within 30° of PB; cleavage passed through this plane; II, male and female pronuclei displaced from PB by 30–90°; cleavage through plane of pronuclei; III, pronuclei aligned within 30° of PB; cleavage displaced from this plane by 30–90°; IV, pronuclei not aligned with PB; cleavage within 30° of the PB. **c**, Zygotes fixed and stained

to reveal maternal (CpG methylated chromatin, pink) and paternal chromatin (blue). Left, prometaphase zygote; right, optical sections through separate two-cell blastomeres. Parental chromosomes do not mix during the first cleavage and still occupy distinct territories at the two-cell stage. **d**, DIC time-lapse series of zygote with cleavage at the site of the PB. **e**, Time-lapse series of H2B–EGFP (green) labelled zygote with red fluorescent bead at sperm entry site. **a**, **c–e**, The diameter of the zygote is approximately 80 μm. **f**, Proportions of zygotes with cleavage plane within 30° (1), 30–60° (2) and 60–90° (3) of the PB in a DIC time-lapse series of 64, a series of 32 to follow H2B–EGFP, and a series of 28 to assess cleavage in relation to the fluorescent bead marking sperm entry.

60° and 90° of the PB (Fig. 1a, f). In those zygotes in which it was located in a medial position (25%) the PB was usually drawn into the furrow as it subsequently ingressed in more than 90% of embryos. Thus, the first cleavage tended to pass close to the PB region and through the plane along which the parental genomes had aligned. We also followed the fluorescence of both GFP-tagged H2B and a bead marking the sperm entry point in time-lapse studies (Fig. 1e and Supplementary Movie 2). This showed that cleavage passed within 30° of the sperm entry position in a significant majority of zygotes (Fig. 1e, f). In addition we performed a small

series of transplantation experiments to re-site the animal pole to observe by time-lapse imaging whether cleavage would then tend to pass through this new site. In almost all (six of seven) embryos the first cleavage did occur within 30° of the newly positioned PB (Supplementary Movie 3). Taken together, these studies of living embryos support and extend earlier work that showed there is a bias for the first cleavage to pass close to the attached PB^{7–9} and site of sperm entry^{1,9}. Recent studies have indicated that cleavage of the two-cell blastomere that inherits the PB also tends to pass close to the PB (ref. 5).

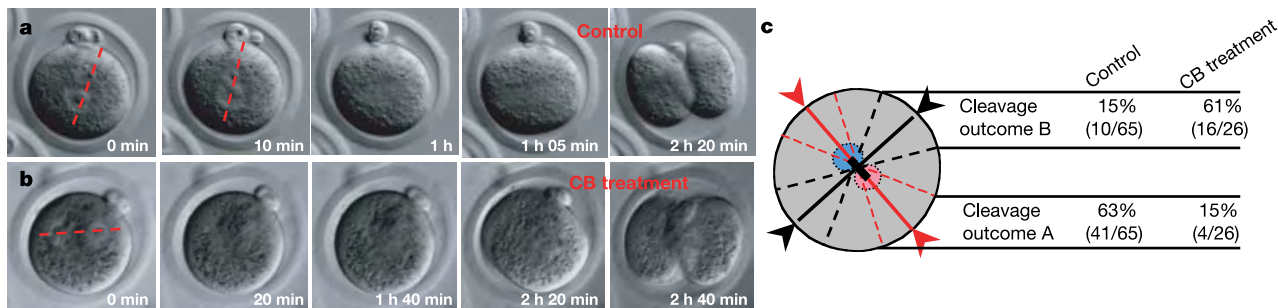


Figure 2 Cytochalasin treatment before pronuclear migration changes the orientation of the first cleavage division. **a**, **b**, Time-lapse series of control (**a**) and cytochalasin-treated (**b**) zygotes. The diameter of the zygotes is approximately 80 μm. **c**, Diagram to show the plane represented by extrapolation of final trajectories of two pronuclei (solid red line).

Cleavage within 30° of this line is represented as ‘outcome A’. The solid black line represents cleavage 90° to this final path of pronuclear migration. Cleavage within 30° of this is referred to as ‘outcome B’. The relative proportions of cleavages that followed these two outcomes in control and cytochalasin-treated embryos are shown.

In the above experiments we neither selected nor excluded embryos, but instead analysed 96% of all recorded zygotes on multiple focal planes at each time point. In contrast, a recent study that examined cleavage with respect to the PB on a single focal plane discarded over 50% of zygotes in which cleavage was not in a suitable position to be observed¹³. Furthermore, to examine cleavage with respect to pronuclei, these same authors selected

embryos in which the pronuclei lay on the same focal plane¹³. This indicates that their data might have been biased through the selection of specific groups of embryos for analysis. It could explain why their conclusion that the first cleavage does not relate to the PB and moreover coincides with a plane between apposing pronuclei differs from the four-dimensional studies described in this paper. Such a bias might arise because in our experience zygotes position

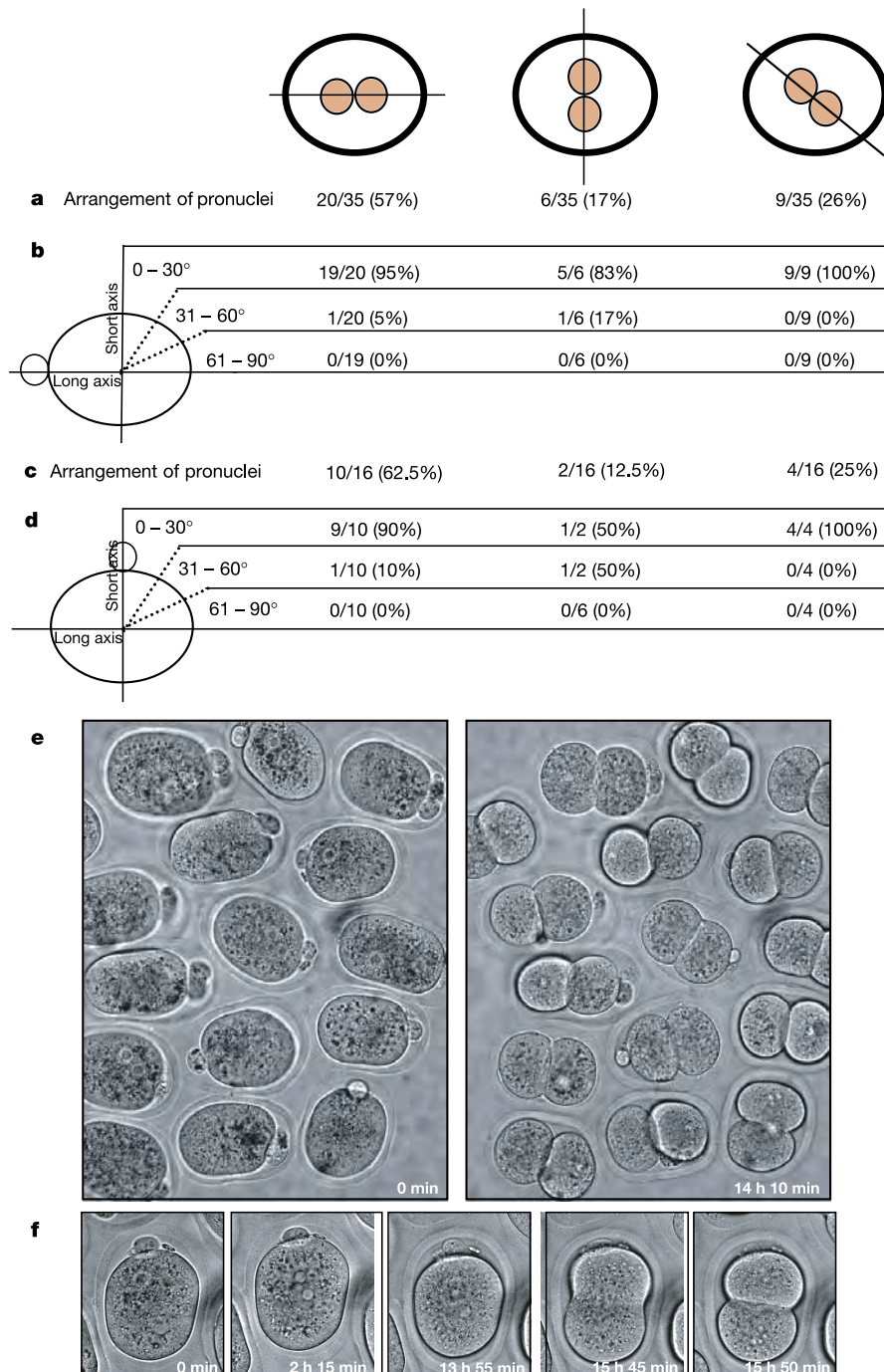


Figure 3 Relationship between final position of pronuclei, long axis of zygote and first cleavage in experimentally elongated zygotes. **a, b**, PB at the end of the long axis; **c, d**, PB at the end of the short axis. **a, c**, Proportions of zygotes in which alignment of pronuclei was within 30° (left), 61–90° (middle) and 31–60° (right) of the long axis in a time-lapse series of 51 zygotes. **b, d**, Proportions of zygotes that showed a cleavage plane within 30° (top), 31–60° (middle) and 61–90° (bottom) of the short axis. **e**, Example of a field of experimentally elongated embryos at the beginning (20 h after hCG injection, left) and end

(34 h 10 min after hCG injection, right) of a time-lapse series. **f**, One example of a zygote at five time points. The ratio between the short and long axis of this zygote was 0.83. Time points: at 0 min (beginning of the experiment 18 h after hCG injection) both pronuclei are still apart from each other; at 2 h 15 min, both pronuclei have met in the centre of the zygote; at 13 h 55 min, the last frame in which pronuclei can still be seen; at 15 h 45 min, elongated zygotes are preparing for cleavage; at 15 h 50 min, cytokinesis occurs. The diameter of the not elongated embryos is approximately 80 μ m.

themselves in culture dishes according to their shape. This is particularly important because zygote shape can override any influence of the region marked by the PB in relation to the orientation of the first cleavage⁹. Could it therefore be that if we were to select zygotes of a particular shape we could also observe a different relationship of the first cleavage to the position of pronuclei? If so, would circumstances in which cleavage followed a different set of cues allow us to test its independent relationship to development of the embryonic–abembryonic axis?

Treatment of zygotes with cytochalasin at concentrations that disrupt the actin cytoskeleton can interfere with the effect of the natural shape change induced by fertilization to affect positioning of the cleavage furrow⁹. We therefore examined cleavage in embryos treated with cytochalasin for 4 h at concentrations that depolymerize the actin cytoskeleton about 6 h before cleavage. In contrast with control zygotes (Fig. 2, outcome A; Supplementary Movie 4), the plane of cleavage in cytochalasin-treated zygotes (Fig. 2, outcome B; Supplementary Movie 5) lay within 30° of a plane orthogonal to the final trajectory of the pronuclei. Thus, only in drug-treated embryos do we see an outcome similar to that reported previously¹³. Indeed, we note that in these experiments, which led to the conclusion that the cleavage plane would lie between apposing transplanted pronuclei¹³, zygotes must have been treated with cytochalasin to enable the necessary micromanipulations. Such a treatment would in itself be enough to influence cleavage orientation. We find that the time of cytochalasin treatment is critical to the outcome. When applied early in the course of pronuclear migration, as is necessary in pronuclear transplantation, cleavage no longer respected the position of the PB (44% cleaved within 30° of the PB, 36% between 30° and 60°, and 20% between 60° and 90°; $n = 25$). If zygotes were treated with cytochalasin once the pronuclei had migrated to the interior, as in our own experiments to transplant cytoplasm from the PB region, most cleavages (73%, 16 of 22) respected the PB.

Next we looked further into the effects of cell shape upon the first cleavage. Our previous studies showed that when the zygote was experimentally ‘flattened’, cleavage passed through the new short axis, which thus overrides the effects of the animal pole and sperm entry site⁹. Could such a shape change influence the positions of pronuclei with respect to the long and short axes of the zygote and their relationship to the cleavage plane? To address this question we changed the shape of zygotes by drawing them into a constraining micropipette and releasing them into sodium alginate such that the PB lay at one end of either the long or the short axis. We found that, independently of the PB position, pronuclei tended to align with the new long axis of the zygote (Fig. 3). In most zygotes (59%, 30 of 51) pronuclei were aligned within 30° of the long axis and only in 16% (8 of 51) were pronuclei within 30° of the short axis. Cleavage occurred within 30° of the short axis in 92% (47 of 51) of all zygotes. Thus, the division plane happens to lie on a plane between the two apposing pronuclei in zygotes whose shape overrides the influence of the PB marked region in positioning cleavage. This contrasts with the great majority of non-manipulated zygotes, in which the pronuclei align with the first cleavage plane. The positions of the pronuclei seem to be independent of the plane of cleavage because in those rare circumstances in which the pronuclei of experimentally elongated zygotes were aligned with the short rather than the long axis (16%, 8 of 51), cleavage still tended to pass through the short axis (Fig. 3). Taken together, these results suggest that it is cell shape rather than position of the pronuclei that influences the cleavage plane in these elongated embryos. Thus, we cannot confirm the previous conclusion¹³ that zygotes always divide according to the plane separating the two apposing pronuclei. We saw such patterns of cleavage only within selected groups of embryos that either had been treated with cytochalasin before the migration of pronuclei or had undergone a shape change such that the pronuclei lay along the long axis.

This ability to manipulate zygotes to divide reproducibly with

this alternative pattern allowed us to determine its consequences for the development of the embryonic–abembryonic axis of the blastocyst. We therefore changed the zygote shape as above and cultured embryos in medium containing sodium alginate until the two-cell stage. Embryos that divided with the PB at one end of the long axis were then lifted from the solidified medium and their blastomeres were labelled with different lipophilic dyes (DiI or DiD, Molecular Probes) applied externally. We allowed such labelled embryos to develop to blastocysts and scored the positions of the clones with respect to the blastocyst cavity by confocal sectioning (Fig. 4). In zygotes whose shape had not been changed, the two respective clones should tend to occupy embryonic (polar trophectoderm and the deeper cells of the inner cell mass) and abembryonic (mural trophectoderm) parts of the blastocyst. The

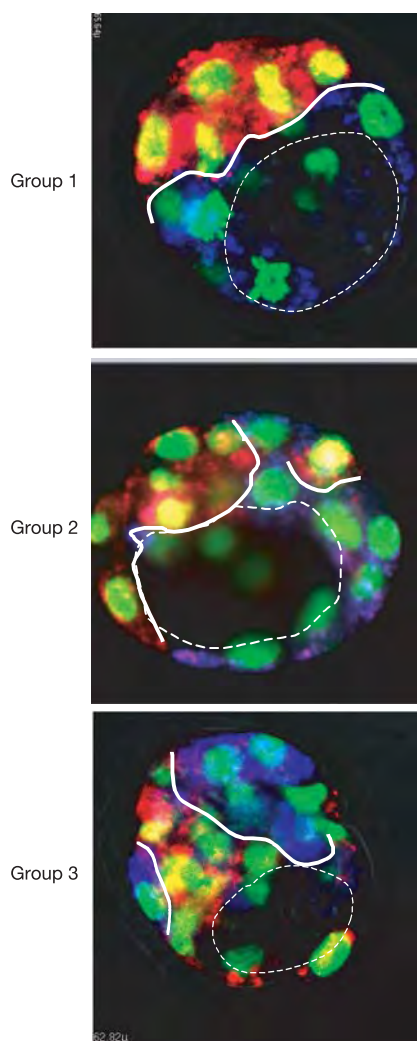


Figure 4 Outcome of changing orientation of first cleavage on the orientation of blastocyst embryonic–abembryonic axis. Zygotes were elongated such that PB was at one end of the long axis. On their division to the two-cell stage, one blastomere was labelled with DiI (red) and other with DiD (blue). At the blastocyst stage, a total of 20 embryos were obtained and classified according to the extent to which cells derived from one blastomere comprising mainly the embryonic part crossed the embryonic–abembryonic boundary zone (a region approximately one cell deep and parallel to the roof of the blastocyst cavity) into the abembryonic part and vice versa. In group 1, up to two cells crossed the boundary zone; in group 2, three cells crossed the boundary zone; and in group 3, more than three cells crossed the boundary zone. Micrographs represent individual optical sections midway through the embryo that show the cavity (thin dashed line) and clonal boundary zone (thick solid line). The diameter of the blastocysts is approximately 80 μ m.

number of cells from each clone in a boundary zone, lying one cell deep and parallel to the blastocoelic surface³, can vary. To compare the distribution of progeny of both two-cell blastomeres between either embryonic or abembryonic parts in relation to previously published studies, we scored the extent to which clones spread beyond the boundary zone in 20 labelled blastocysts manipulated so that the PB had been at one extreme of the long axis after the first cleavage (Fig. 4 and Supplementary Fig. 1). If either no or a few cells (up to two, about 3–9% of the total mean cell number ($n = 22.25$) at this stage) had spread beyond the boundary zone, blastocysts were scored in group 1; if three cells crossed, blastocysts were scored in group 2; and if more than three, blastocysts were scored in group 3. In group 1, in 85% of blastocysts (17 of 20) two or fewer cells from predominantly embryonic clones were positioned beyond the boundary zone in the abembryonic part (Fig. 4), and in 65% (13 of 20) of the same blastocysts, two or fewer cells from predominantly abembryonic clones were located in the embryonic part (Fig. 4). In group 2, in 15% (3 of 20) of blastocysts three cells from the predominantly embryonic clone were located in the abembryonic part; in the other direction, the proportion was 20% (4 of 20 blastocysts). In group 3 there were no (0 of 20) blastocysts in which three or more cells from the predominantly embryonic clone were present in the abembryonic part. Only 15% (3 of 20) of blastocysts in this group showed the predominantly abembryonic clone extending into the embryonic part. Thus, in most embryos, most progeny of two-cell blastomeres came to lie in either the embryonic or abembryonic part of the blastocyst. A similar proportion of embryos showed this relationship in previous studies of non-manipulated embryos in which cleavage tended to pass close to the animal pole³. We also observed relatively little cell mixing between the two clones at the early blastocyst stage, in agreement with previous reports¹⁴ (Supplementary Fig. 1).

Our study adds support to earlier evidence for a relationship between the first cleavage and a region close to PB^{5–9}. However, this relationship of the animal pole region to cleavage orientation can be overridden by the effects of the shape upon the zygote as we have demonstrated previously⁹. We now show that, irrespective of the orientation of cleavage in relation to the pronuclei or the PB, in most embryos one two-cell blastomere populates the embryonic part of the blastocyst, and its sister populates the abembryonic part. Thus, formation of the embryonic–abembryonic axis tends to occur with respect to the first cleavage plane itself whether or not it passes close to the animal pole. Does this mean that the organization of the egg in relation to the animal pole has no influence over development other than normally coordinating the first cleavage? Recent experiments suggest that the patterns of early cleavage divisions segregate blastomeres not only with different fates but also with differing developmental properties^{5,6}. This relates both to their specific position in the embryo with respect to the animal pole and the differential timing of the division of these cells. Thus, the existence of spatially distributed components of the egg working in concert with factors regulating cell division to influence development cannot be discounted. Further studies are required to identify any such putative components and to establish whether changing the orientation of the first cleavage might lead to their repositioning and indeed to the altered developmental properties of four-cell-stage blastomeres. □

Methods

Zygotes were collected from F₁ (C57BL/6 × CBA) females induced to superovulate as described previously³ and then mated with males of the same strain. In experiments in which chromatin was detected by fluorescence, zygotes were collected from CAG:H2B–EGFP¹⁰ transgenic mice mated with males of the same strain. Zygotes were collected 23 h after injection with human chorionic gonadotrophin (hCG) as described previously³. In experiments examining the significance of the actin cytoskeleton in the subsequent process of spindle orientation, we treated zygotes with 1 or 5 µg ml^{−1} cytochalasin B for 3–4 h starting 18–19 h after injection with hCG.

Time-lapse imaging was performed with either a Nikon inverted microscope with a

Princeton Instruments camera using IP lab software or with a Leica inverted microscope with a Hamamatsu Orca camera using Openlab software (Improvision). Embryos were placed in 0.8 ml of M2 medium under paraffin oil in a glass-bottomed dish on a heating stage at 37 °C. Images were recorded at 5-min intervals for up to 12 h. The positions of pronuclei were measured with Velocity software (Improvision) and described as the coordinates of the centre of each pronucleus (x_1, y_1, z_1 and x_2, y_2, z_2). To define the division plane we used the coordinates of the centres of two blastomeres by the end of cytokinesis (x_3, y_3, z_3 and x_4, y_4, z_4). The angle between division plane and pronuclear position was then calculated from the equation

$$\theta = \arccos \frac{((x_2 - x_1)(x_4 - x_3) + (y_2 - y_1)(y_4 - y_3) + (z_2 - z_1)(z_4 - z_3))}{\sqrt{((x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2)((x_4 - x_3)^2 + (y_4 - y_3)^2 + (z_4 - z_3)^2)}}$$

To label the sperm entry point the procedure was as described previously¹, except that red rather than green fluorescent beads were used so that they could be distinguished from H2B–GFP. Animal pole transplantation was performed as described previously⁸. Zygotes were treated with 1 µg ml^{−1} cytochalasin D for 30–40 min immediately before transplantation.

Immunostaining was performed on embryos fixed with 3.7% paraformaldehyde in phosphate-buffered saline and processed as previously⁸. The shapes of zygotes were changed and their resulting dimensions measured as described previously⁹. Zygotes were elongated 18–20 h after hCG injection such that the mean ratio between the short and the long axis was 0.7 (range 0.6–0.9).

In experiments to monitor the effect of changing the position of the first cleavage on the development of the embryonic–abembryonic axis, zygotes were elongated 25–27 h after hCG injection. They were subsequently cultured in medium containing sodium alginate for the next 17–18 h and then lifted from it when they divided to the two-cell stage. Both blastomeres of these embryos were subsequently labelled: one blastomere with DiI, the other with DiI. The labelling and subsequent culture was performed as described previously³. Lineage tracing experiments were performed on the H2B–GFP transgenic line to facilitate analysis of the distribution of cells at the blastocyst stage. Blastocysts were observed by confocal microscopy and images were captured at 7-µm intervals. By examining all sections (nine or ten) in each series it was possible to determine the distribution of cells labelled by specific dyes in the embryonic (polar trophectoderm and deeper inner-cell mass cells) and abembryonic (mural trophectoderm) parts of the blastocyst. The boundary zone between these two parts was defined as a cell layer about one cell deep and parallel to the roof of the blastocoel cavity, as previously³.

Received 9 November 2004; accepted 25 January 2005; doi:10.1038/nature03388.

- Piotrowska, K. & Zernicka-Goetz, M. Role for sperm in spatial patterning of early mouse embryos. *Nature* **409**, 517–521 (2001).
- Gardner, R. L. Specification of embryonic axes begins before cleavage in normal mouse development. *Development* **128**, 839–847 (2001).
- Piotrowska, K., Wianny, F., Pedersen, R. A. & Zernicka-Goetz, M. Blastomeres arising from the first cleavage division have distinguishable fates in normal mouse development. *Development* **128**, 3739–3748 (2001).
- Fujimori, T., Kurotaki, Y., Miyazaki, J. I. & Nabeshima, Y. I. Analysis of cell lineage in 2- and 4-cell mouse embryos. *Development* **21**, 5113–5122 (2003).
- Piotrowska, K., Zernicka-Goetz, M. Spatial arrangement of individual 4-cell stage blastomeres and the order in which they are generated correlate with blastocyst pattern in the mouse embryo. *Mech. Dev.* advance online publication, 18 December 2004 (doi:10.1016/j.mod.2004.11.014).
- Piotrowska, K., Perea-Gomez, A., Haraguchi, S. & Zernicka-Goetz, M. Four-cell stage mouse blastomeres have different developmental properties. *Development* **132**, 479–490 (2005).
- Gardner, R. L. The early blastocyst is bilaterally symmetrical and its axis of symmetry is aligned with the animal–vegetal axis of the zygote in the mouse. *Development* **124**, 289–301 (1997).
- Plusa, B., Grabarek, J. B., Piotrowska, K., Glover, D. M. & Zernicka-Goetz, M. Site of the previous meiotic division defines cleavage orientation in the mouse embryo. *Nature Cell Biol.* **4**, 811–815 (2002).
- Gray, D. et al. First cleavage of the mouse embryos responds to change in egg shape at fertilisation. *Curr. Biol.* **14**, 397–405 (2004).
- Hadjantonakis, A.-K. & Papaioannou, V. High resolution dynamic *in vivo* imaging and tracking in mice. *BMC Biotechnol.* **4**, 33 (2004).
- Mayer, W., Smith, A., Fundele, R. & Haaf, T. Spatial separation of parental genomes in preimplantation mouse embryos. *J. Cell Biol.* **148**, 629–634 (2000).
- Donahue, R. P. Fertilization of the mouse oocyte: sequence and timing of nuclear progression to the two-cell stage. *J. Exp. Zool.* **180**, 305–318 (1972).
- Hiragi, T. & Solter, D. First cleavage plane of the mouse egg is not predetermined but defined by the topology of the two apposing pronuclei. *Nature* **430**, 360–364 (2004).
- Garner, W. & McLaren, A. Cell distribution in chimaeric mouse embryos before implantation. *J. Embryol. Exp. Morphol.* **32**, 495–503 (1974).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This work was supported by a Wellcome Trust Senior Research Fellowship to M.Z.-G. and a BBSRC Project Grant to M.Z.-G. and D.M.G. K.P.N. was holding a Marie Curie Fellowship from the European Union. V.E.P. acknowledges support from the NIH. B.P. is on leave from the Department of Experimental Embryology at The Polish Academy of Science, Jastrzebiec, Poland.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.Z.-G. (mzg@mole.bio.cam.ac.uk).

The *MET* oncogene drives a genetic programme linking cancer to haemostasis

Carla Boccaccio, Gabriella Sabatino, Enzo Medico, Flavia Girolami, Antonia Follenzi, Gigliola Reato, Antonino Sottile, Luigi Naldini & Paolo M. Comoglio

Division of Molecular Oncology, Institute for Cancer Research and Treatment, University of Turin Medical School, Str. Prov. 142, I-10060 Candiolo, Torino, Italy

The close relationship between activation of blood coagulation and cancer is an old enigma. In 1865, *migrans thrombophlebitis* ('a condition of the blood that predisposes it to spontaneous coagulation') was described as a forewarning of occult malignancy (Trousseau's sign¹). This pioneering observation emphasized the existence of haemostasis disorders associated with cancer onset; this phenomenon has since been extensively reported in clinical and epidemiological studies^{2–4}, but has so far resisted a mechanistic explanation. Here we report a mouse model of sporadic tumorigenesis based on genetic manipulation of somatic cells. Targeting the activated, human *MET* oncogene to adult liver caused slowly progressing hepatocarcinogenesis. This was preceded and accompanied by a syndrome manifesting first with blood hypercoagulation (venous thromboses), and then evolving towards fatal internal haemorrhages. The pathogenesis of this syndrome is driven by the transcriptional response to the oncogene, including prominent upregulation of plasminogen activator inhibitor type 1 (PAI-1) and cyclooxygenase-2 (COX-2) genes. *In vivo* analysis showed that both proteins support the thrombohaemorrhagic phenotype, thus providing direct genetic evidence for the long-sought-after link between oncogene activation and haemostasis.

To study the early events of transformation in a setting that recapitulates the natural occurrence of sporadic cancer, it would be ideal to develop a model that initiates from single somatic cells harbouring mutations of one or more cancer-related genes, but intermingled with their normal tissue context. Here we have developed a mouse model of somatic oncogene delivery, relying on the unique properties of late-generation lentiviral vectors⁵. These enable oncogene targeting by non-viral promoters in postnatal life, and transformation of somatic cells that remain functionally interlocked within the normal tissue.

The *MET* oncogene encodes the tyrosine kinase receptor for hepatocyte growth factor/scatter factor (HGF/SF), and drives a physiological cellular programme known as 'invasive growth', which underlies tissue morphogenesis and repair. Aberrant execution of this programme in time and localization has been associated with neoplastic transformation, invasion and metastasis (for review see ref. 6). Activation of the *MET* oncogene has been reported in a wide array of human tumours, either as a consequence of germline or sporadic mutations⁷, or, more commonly, of over-expression⁸, probably caused by hypoxia-induced transcription⁹.

We cloned an activated cytoplasmic form of the human *MET* oncogene (*TPR-MET*; ref. 10) under the hepatocyte-specific albumin (ALB) promoter, into a self-inactivating lentiviral vector (LV)¹¹ (ALB-*MET*-LV). As a control, we generated a lentiviral vector encoding green fluorescent protein¹² (GFP, ALB-GFP-LV). Upon ALB-*MET*-LV transduction, mouse liver progenitor cells (MLP29) underwent a marked phenotypic change, increased their proliferation rate, lost contact inhibition and formed foci of transformation (data not shown). After transplantation into the spleen of immunodeficient mice (C.B-17/1CrCrl-scid or SCID), the oncogene-transduced cells colonized the site of injection and, as expected,

the liver¹³. Within 4 weeks, transplanted cells formed tumours resembling glandular hepatocellular carcinomas and/or hepatocellular carcinomas (data not shown).

We then injected vector preparations of equivalent infectivity (ALB-*MET*-LV or ALB-GFP-LV) into the tail vein of SCID mice. Two weeks later, we detected expression of the transgene in $5 \pm 1.4\%$ of hepatocytes (data not shown). Three months later,

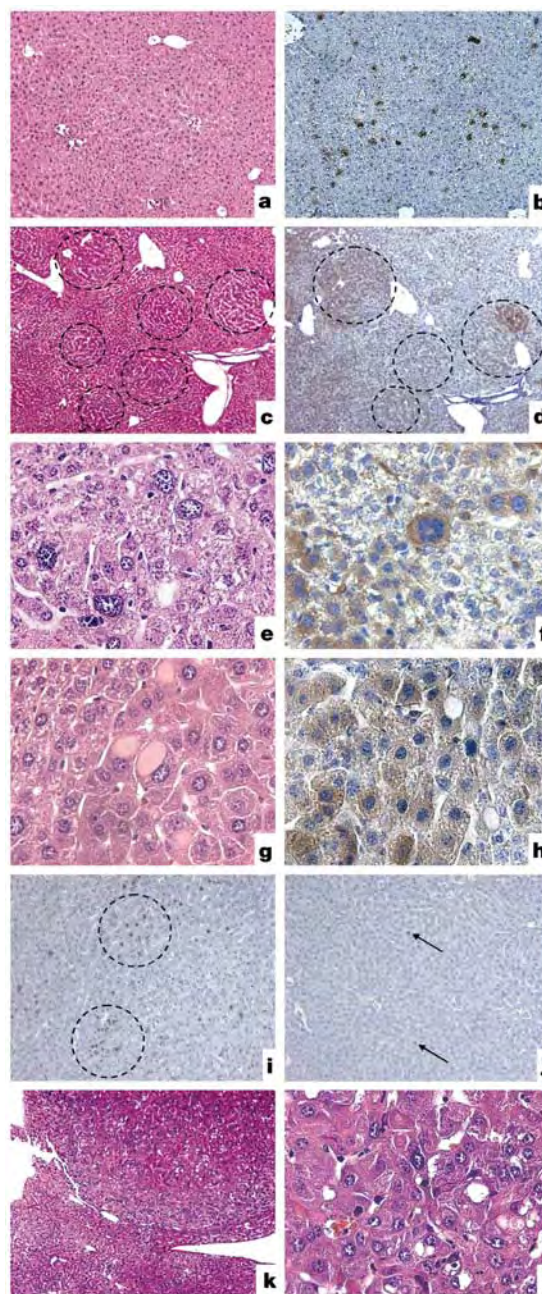


Figure 1 Pre-neoplastic and neoplastic lesions after liver transduction of the *MET* oncogene. **a, b**, Control liver sections three months after GFP transduction. **a**, H&E staining. **b**, Anti-GFP immunohistochemistry. **c, d**, Sequential sections of liver transduced with *MET*, showing several basophilic nodules. **c**, H&E staining. **d**, Anti-*MET* immunohistochemistry. **e–h**, Details of cellular alterations in the nodules: nuclear atypia (H&E staining in **e**, anti-*MET* immunohistochemistry in **f**) and cytoplasmic inclusions (H&E staining in **g**, anti-*MET* immunohistochemistry in **h**). **i, j**, Mitotic cells stained by anti-PCNA in nodules of *MET*-transduced livers (**i**) or in GFP controls (arrows in **j**). **k, l**, Liver carcinoma six months after *MET* transduction (H&E staining, **k**) showing high-grade atypia (H&E staining, **l**).

all mice ($n = 10$ of the 10 killed) transduced with the activated *MET* oncogene showed multiple basophilic nodules (Fig. 1c, d) of high-grade dysplastic hepatocytes, with nuclear and cytoplasmic alterations (Fig. 1e–h) and high proliferation rates (Fig. 1i, j). Immunohistochemistry showed that nodules of dysplastic hepatocytes were invariably and homogeneously positive for expression of the *MET* oncogene (Fig. 1d, f, h). This strongly indicates that each nodule was derived from clonal expansion of single transduced hepatocytes. No pathological alterations were observed in livers of mice ($n > 50$) transduced with the same vector driving GFP expression^{12,14} (Fig. 1a, b). In the 5 (out of 25) mice that survived the thrombohaemorrhagic syndrome (described below) for six months, a few nodules progressed to overt neoplasia (Fig. 1k, l).

Three to four weeks after vector injection, and before the appearance of dysplastic nodules in the liver (at three months), all mice ($n = 25$) infected with ALB-*MET*-LV, but none of an equal number of GFP controls, developed a thrombohaemorrhagic syndrome. This began with formation of venous thrombi visible through the glabrous tail skin (Fig. 2a); these thrombi then underwent progressive organization (Fig. 2b). Within two months of *MET* transduction, intravascular coagulation invariably evolved into a progressive haemorrhagic diathesis, which was lethal for the mice within 6–7 months. First, petechiae (small subcutaneous haemorrhages) appeared under the glabrous skin of the extremities; then, an extended haematoma appeared in the lower back and/or internal haemorrhages occurred, which, at necropsy, were mostly found in the abdomen as retroperitoneal or periovaric haematomas (Fig. 2c, d). This chronic association (over weeks or months) of intravascular coagulation with bleeding complications is characteristic of chronic disseminated intravascular coagulation (DIC) syndrome, which has been described in a very large number of cancer patients^{15,16}. Analysis of the mouse peripheral blood coagulation system supported the diagnosis of DIC (Table 1): platelet number progressively decreased during the evolution of the syndrome, as a sign of their recruitment into clot formation; production of the fibrin digestion fragment ‘D-dimer’ progressively increased, as a

sign of clot degradation; and activated partial thromboplastin time and prothrombin time values were significantly prolonged ($P < 0.05$) in the haemorrhagic phase, a typical sign of ‘consumption coagulopathy’ (that is, depletion of coagulation factors, which impairs blood clotting). Altogether, these data were consistent with the dynamic nature of DIC syndrome, which involves hyperactivation of the coagulation system and continuous fibrin deposition and degradation, and ultimately ends in haemostasis failure and fatal haemorrhages¹⁶. Moreover, in *MET*-transduced mice, chronic hyperactivation of the haemostatic system, with massive platelet recruitment and destruction, stimulated compensatory extramedullary haemopoiesis in the spleen, leading to organ enlargement and enrichment in haemopoietic precursors. Megakaryocyte levels, in particular, were consistently elevated in the bone marrow and in the spleen (Fig. 2e, f).

The observed haemorrhages were not a consequence of hepatic insufficiency, as serum levels of liver enzymes were normal (data not shown). As hepatocytes are the synthesis site of several coagulation factors, we investigated whether the activated *MET* oncogene could drive the above haemostasis disorders in mice transplanted with mammary epithelial cells (MDA-MB-435) that had been transduced with CMV-*MET*-LV (a lentiviral vector driving *MET* expression through the constitutive human cytomegalovirus (CMV) promoter). A lethal DIC syndrome was invariably observed about four weeks after spleen injection, and was preceded by a consistent alteration of coagulation parameters. Platelets decreased from $935 \pm 44 \times 10^3 \text{ mm}^{-3}$ to $330 \pm 52 \times 10^3 \text{ mm}^{-3}$ ($P < 0.05$), and partial thromboplastin time increased from $28 \pm 2 \text{ s}$ to $40 \pm 15 \text{ s}$ ($P < 0.05$).

The induction of haemostasis disorders by *MET* indicates a direct effect on gene transcription. Through genome-wide expression profiling of MLP29 cells transduced with the activated *MET* oncogene, we observed that many of the 71 haemostasis genes represented in the microarrays (12,000 genes explored) were slightly upregulated (Supplementary Table 1). Two genes were greatly induced (Fig. 3): serine proteinase inhibitor clade E member 1

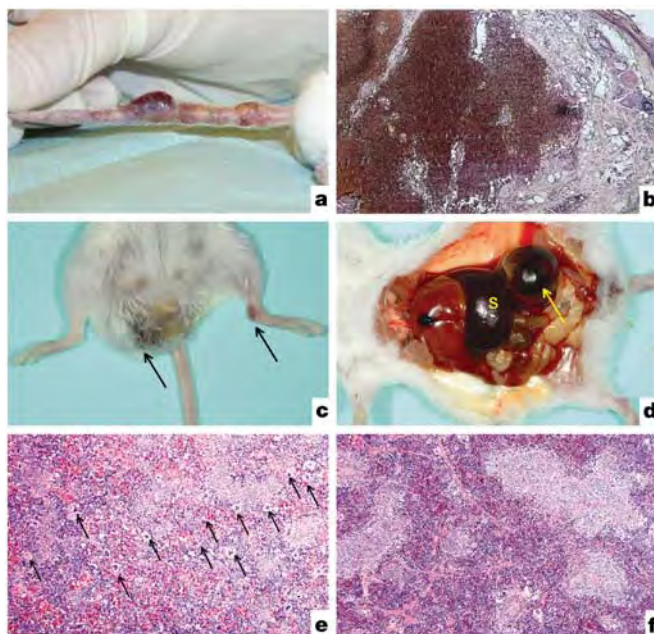


Figure 2 The thrombohaemorrhagic syndrome induced by transduction of the *MET* oncogene. **a**, Venous thrombosis in the mouse tail. **b**, Thrombus section (H&E staining). **c**, **d**, Haemorrhages in the subcutis (arrows in **c**) or in the ovary (arrow in **d**). The spleen (S) is also enlarged. **e**, Hypercellularity with megakaryocytes (arrows) in spleen section (H&E staining) from *MET*-transduced mice. **f**, Spleen section of GFP control mice (H&E staining).

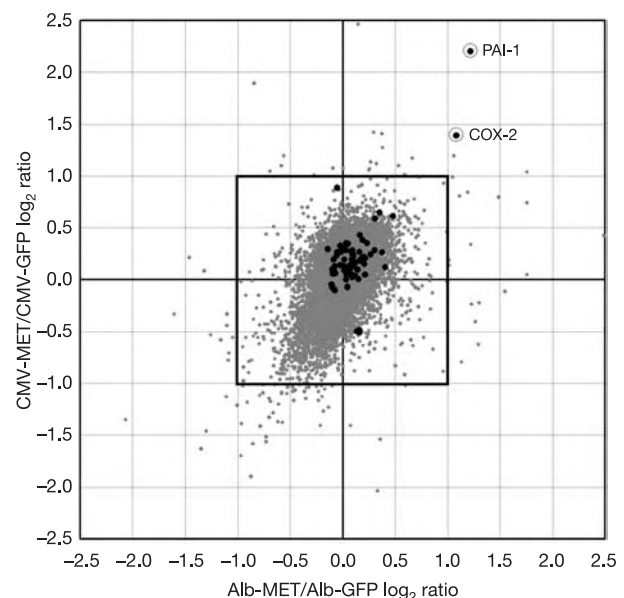


Figure 3 Upregulation of haemostasis genes in MLP29 hepatocytes after transduction of the *MET* oncogene. Comparison of transcriptional changes induced by *MET* under ALB or CMV promoter in MLP29 with respect to control cells transduced with GFP. Log₂ ratios between cells transduced with *MET* or GFP under the albumin promoter (x axis); log₂ ratios between cells transduced with *MET* or GFP under the CMV promoter (y axis). Black dots, genes involved in haemostasis; grey dots, all other genes. The threshold of twofold change is highlighted as a box delimiting a log₂ ratio of ± 1 .

(better known as plasminogen activator inhibitor type 1, PAI-1) and prostaglandin-endoperoxide synthase 2 (also called cyclooxygenase-2, COX-2). PAI-1 was the most heavily upregulated gene, not only within the haemostasis gene subset, but also in the whole 12,000 gene set, and COX-2 was the third most-induced gene (a complete analysis of the whole data set will be reported elsewhere). Upregulation of PAI-1 and COX-2 proteins was confirmed by western blot analysis (data not shown). Interestingly, COX-2 and PAI-1 expression was also constitutively induced in mammary epithelial cells (MDA-MB-435) transduced by CMV-MET-LV (data not shown), and transiently increased in normal hepatocytes (MLP29) 1 h after stimulation with HGF/SF, the physiological ligand for the wild-type MET receptor (Supplementary Fig. 1).

PAI-1 is a well-known inhibitor of fibrin degradation¹⁷; COX-2 is responsible for the synthesis of prostaglandins, the derivatives of which modulate platelet functions¹⁸. We thus investigated whether mice expressing the activated *MET* transgene and developing the thrombohaemorrhagic syndrome released PAI-1 protein and COX-2 metabolic products in plasma and in urine, respectively. In tandem with the onset of thrombosis (three weeks after oncogene transduction), the plasma level of PAI-1 was increased threefold with respect to GFP-transduced mice (Fig. 4a). At the same time, we found increased levels of prostacyclin catabolites in urine (Fig. 4b). Prostacyclin is a major downstream product of COX-2 activation. PAI-1 and COX-2 proteins were specifically localized to *MET*-transduced hepatocytes, and were evident in dysplastic nodules, as shown by immunohistochemistry (Fig. 4c–f). These data strongly indicate that PAI-1 and COX-2 can systemically sustain the *MET*-associated thrombohaemorrhagic syndrome. To assess their causal role in the haemostasis disorders, we administered XR5118, a PAI-1 inhibitor¹⁹, or Rofecoxib, a COX-2 inhibitor²⁰, to *MET*-transduced mice. In the presence of either drug, the evolution of the syndrome towards full-blown DIC was inhibited, as shown by monitoring the appearance of the circulating fibrin degradation product D-dimer¹⁶ (Table 2). To identify any interdependence between tumour development and coagulopathy, we also investigated whether COX-2 inhibition affects the progression of pre-neoplastic nodules (for example, by hampering ‘tumour nesting’ through inhibition of the ‘provisional fibrin matrix’²¹). At early time points after *MET* transduction (1–3 months) no significant effects

on nodule development were observed; but later on (4–6 months) clear and widespread signs of nodule degeneration appeared, consisting of massive areas of necrosis, surrounded by scattered signs of apoptosis and inflammatory reactions (Fig. 4g, h). This observation is in line with previous results showing that COX-2 inhibition suppresses HGF/*MET*-fostered cancer cell survival²², and also is in line with the finding that COX-2 inhibitors suppress progression of intestinal polyposis both in mouse models and in human patients²³.

The observed association of a thrombohaemorrhagic syndrome with *MET*-induced carcinogenesis provides the first animal model for studying the prothrombotic state of cancer patients. In this model, we have established a link between oncogene activation, cell transformation and haemostasis disorders. Previous evidence has implicated the two major transcriptional targets of *MET*, PAI-1 and

Table 1 Blood coagulation in mice transduced with the *MET* oncogene

Transgene	Parameter	Time after transduction (days)		
		0	30	90
GFP	Platelets ($\times 10^3 \text{ mm}^{-3}$)	968.5 $\pm$ 76.2	655.7 $\pm$ 85.8	800 $\pm$ 50
	D-dimer ($\mu\text{g ml}^{-1}$)	<0.05	<0.05	<0.05
	PT (s)	12.4 $\pm$ 0.5	11.6 $\pm$ 0.28	11.4 $\pm$ 0.6
	PTT (s)	28.5 $\pm$ 2.5	28.3 $\pm$ 1.38	27.3 $\pm$ 0.5
	Platelets ($\times 10^3 \text{ mm}^{-3}$)	973.8 $\pm$ 45.7	350.3 $\pm$ 113.5*	149.4 $\pm$ 41†
<i>MET</i>	D-dimer ($\mu\text{g ml}^{-1}$)	<0.05	0.11 $\pm$ 0.2	0.22 $\pm$ 0.1
	PT (s)	12.9 $\pm$ 0.4	11.8 $\pm$ 1.1	25.1 $\pm$ 4.25
	PTT (s)	29 $\pm$ 2.6	24.1 $\pm$ 1.15	32.7 $\pm$ 0.35

PTT, partial thromboplastin time; PT, prothrombin time.

*Onset of the thrombotic phase.

†Onset of the haemorrhagic phase.

Table 2 D-dimer ($\mu\text{g ml}^{-1}$) in mice transduced with the *MET* oncogene

Transgene	Inhibitor	Time after transduction (days)	
		0	60
GFP	Control	<0.05	<0.05
	Rofecoxib	<0.05	<0.05
	XR5118	<0.05	<0.05
	Control	<0.05	0.21 $\pm$ 0.02
<i>MET</i>	Rofecoxib	<0.05	0.12 $\pm$ 0.01
	XR5118	<0.05	0.13 $\pm$ 0.02

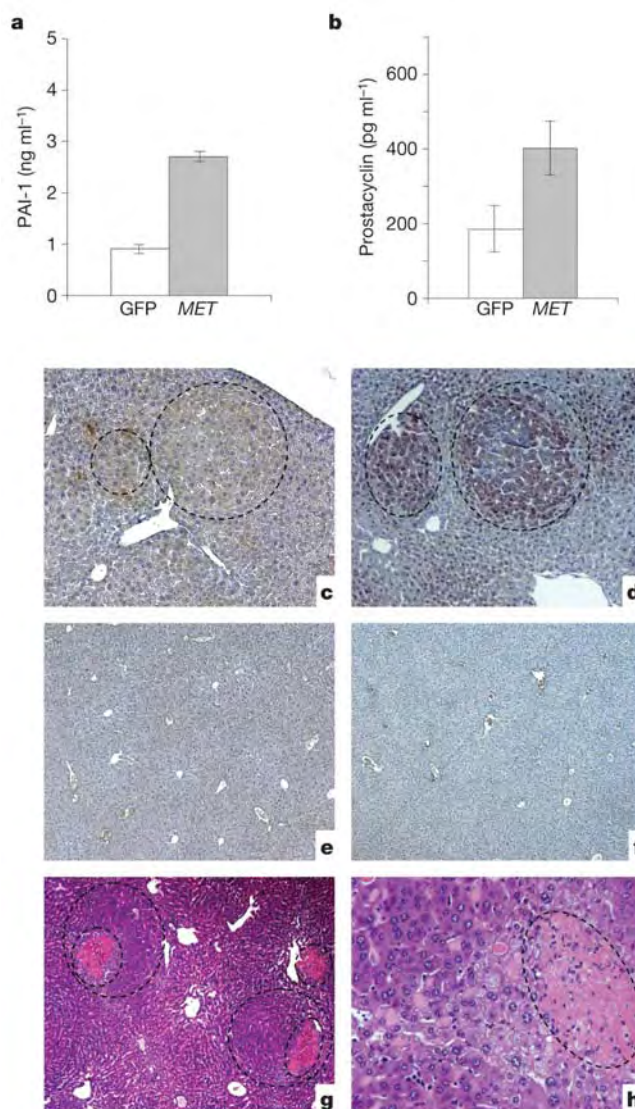


Figure 4 Expression of PAI-1 and COX-2 in *MET*-transduced mice. **a, b**, Plasma levels of PAI-1 (**a**) and urinary levels of catabolites (6-keto-prostaglandin $F_{1\alpha}$ and 2,3-dinor-6-keto-prostaglandin $F_{1\alpha}$) of the COX-2 product prostacyclin (**b**), both measured three weeks after transduction with ALB-MET-LV (*MET*) or ALB-GFP-LV (GFP). Bars indicate s.e.m.; $P < 0.05$. **c, d**, Expression of PAI-1 (**c**) and COX-2 (**d**) protein in pre-neoplastic nodules, detected by immunostaining with specific antibodies. **e, f**, Basal level of PAI-1 (**e**) and of COX-2 expression (**f**), measured in GFP-transduced livers. **g, h**, Necrotic areas in nodules from mice treated with the COX-2 inhibitor Rofecoxib (H&E staining; **h** is a higher magnification of **g**).

COX-2, in the control of both haemostasis and cancer progression^{23,24}. The association of the known pro-invasive properties of the *MET* oncogene with the pro-coagulant activity described here indicates that the two functions are evolutionarily related and possibly interdependent. This connection is strengthened by the structure of HGF/SF itself—the ligand for the tyrosine kinase receptor encoded by the *MET* proto-oncogene. HGF/SF is an uncommonly large protein for a growth factor; it belongs to the plasminogen family as it contains 'kringle' motifs and a serine protease-like domain²⁵. Moreover, as with clotting factors, HGF/SF undergoes proteolytic cleavage to become biologically active; this is performed by coagulation enzymes (for a review see ref. 26). It is therefore tempting to speculate that HGF/SF and its receptor *MET* are themselves members of the coagulation cascade, and evolved to modulate cellular functions such as invasive growth⁶, which takes advantage of fibrin polymerization. In tumours as well as in wound healing, the fibrin matrix forms a rudimentary scaffold that is an absolute requirement for supporting the growth of the tumour itself and the incoming new vessels²¹. The data presented here show that oncogenic activation of *MET* triggers a genetic programme that not only transforms cells but also creates an extracellular environment that is fertile for tumour cell expansion and invasion. It has not escaped our attention that pharmacological interference with the above process may be useful for prevention or treatment of invasive cancer. □

Methods

Lentiviral vectors and *in vitro* gene transduction

ALB-GFP-LV and CMV-GFP-LV transfer vectors have already been described as pRRLsin.cPPT.ALB.GFP.Wpre and pRRLsin.sPPT.CMV.GFP.Wpre, respectively¹². ALB-MET-LV and CMV-MET-LV were generated by replacing GFP with a 1,700 base pair (bp) *EcoRI/SalI* fragment containing the human *TPR-MET* sequence (GenBank accession number U19348). High-titre purified vector stocks were produced in 293T cells as described²⁷ by co-transfection of transfer vector with the packaging constructs pMDLg7/pRRE and REV¹³, and the pMD2.G construct encoding the vesicular stomatitis virus (VSV.G) envelope. MLP29 (mouse liver progenitor clone 29, ref. 28) or MDA-MB-435 cells were transduced by adding 360 ng of vector p24 to 3×10^5 cells, as described²⁷.

In vivo gene delivery and cell transplantation

Female C.B-17/ICrCrI-scid (SCID) mice (Charles River Laboratories) were manipulated according to protocols approved by the Turin University Bioethical Committee and the Italian Ministry of Health. Anaesthesia was performed with tiletamine and zolazepam (Zoletil 20, Virbac, $45 \mu\text{g g}^{-1}$ body weight) and xylazine (Rompun, Bayer, $7.5 \mu\text{g g}^{-1}$ body weight). For gene delivery, $20 \mu\text{g}$ of vector p24 was injected into the mouse tail vein. For intrasplenic transplantation, 10^6 MLP29 or MDA-MB-435 cells, transduced as above, were injected into the inferior pole of the spleen.

Pathology and immunohistochemistry

Mice were killed by CO_2 inhalation. Organs and tissues were embedded in paraffin, according to standard procedures. Sections were stained with Mayer's haematoxylin and eosin Y (H&E) or underwent immunohistochemical staining with one of the following primary antibodies: a rabbit anti-human *MET* polyclonal antibody (C12, Santa Cruz Biotechnology, 1:100) specifically recognizing the human *MET* sequence; a mouse monoclonal antibody against proliferating-cell nuclear antigen (PCNA; BD Biosciences Pharmingen, 1:50); a rabbit anti-COX-2 polyclonal antibody (Cayman Chemical Company, 1:50); and a rabbit anti-PAI-1 polyclonal antibody (American Diagnostica, 1:30). Primary antibodies were detected by EnVision + System Peroxidase-DAB (DAKO Cytomation). Direct microscopy micrographs were captured using a Leica DC 300F videocamera and elaborated using Leica IM50 software.

Clinical chemistry

Blood was collected by tail bleeding or cardiac puncture after a sub-lethal dose of anaesthetic and analysed using Abbott Cell-Dyn 3500 for cell count, or STA-Compact (Roche Diagnostic) for prothrombin time, activated partial thromboplastin time and D-dimer measurement. PAI-1 concentration was measured in duplicate on independent plasma samples from three mice using the enzyme-linked immunosorbent assay (ELISA) kit MPAIKT-TOT (Molecular Innovations). Urine samples were collected from pairs of mice during an approximately 4 h stay in a diuresis cage (Tecniplast). The major catabolites of the COX-2 product prostacyclin (6-keto-prostaglandin $\text{F}_{1\alpha}$ and 2,3-dinor-6-keto-prostaglandin $\text{F}_{1\alpha}$) were measured in duplicate on three independent urine samples using Correlate-EIA urinary prostacyclin enzyme immunoassay kit (Assay Designs). All values are expressed as average $\pm$ standard error of the mean. The differences between experimental groups were analysed by Student's *t*-test and considered statistically significant at $P \leq 0.05$.

Gene expression

MLP29 and MDA-MB-435 were transduced as above. MLP29 cells were synchronized by two one-week cycles of serum starvation. Each experimental condition (wild-type cells or cells transduced with ALB-MET-LV, CMV-MET-LV, ALB-GFP-LV or CMV-GFP-LV) was tested in duplicate. Total RNA was isolated and further purified using RNeasy MicroKit (Qiagen), according to the manufacturer's protocols. Biotin-labelled complementary RNAs were obtained from $10 \mu\text{g}$ of total RNA samples according to standard Affymetrix protocols and hybridized to MGU74subA mouse arrays (Affymetrix). Raw data were normalized using DNA-Chip Analyser (dChip)²⁹. Average signal values from control or *MET* points and average $\log_2$ ratios were calculated using Microsoft Excel. A panel of 71 transcripts of genes controlling haemostasis was extracted from the whole data set by keyword searching of gene ontology terms (Mouse Genome Informatics, <http://www.informatics.jax.org>) and gene definitions. The median $\log_2$ ratio between *MET*-transduced cells and controls observed for the whole haemostasis gene panel was tested for significance using Montecarlo analysis. Distribution of 50,000 median $\log_2$ ratios from randomly selected 71-gene panels indicated that the *P*-value for the observed $\log_2$ ratio is $<2.5 \times 10^{-4}$. Transcriptional regulation of PAI-1 and COX-2 by HGF/SF on MLP29 cells was assessed with Affymetrix MGU74subA microarrays, according to standard procedures (the whole data set will be reported elsewhere).

Pharmacology

Starting one week after lentiviral transduction, the PAI-1 inhibitor XR5118 (ref. 19; a gift from P. Charlton and P. Andreasen) was administered intraperitoneally to mice (5 mg kg^{-1} body weight, three times a week); the COX-2 inhibitor Rofecoxib (provided by Merck Research Laboratories) was continuously supplied with mouse chow (0.0075% w/w; ref. 30; Charles River Laboratories).

Received 20 October 2004; accepted 13 January 2005; doi:10.1038/nature03357.

1. Trouseau, A. Phlegmasia alba dolens. *Clinique Médicale de l'Hotel-Dieu de Paris* Vol. 3, 654–712 (J. B. Ballière et Fils, Paris, 1865).
2. Baron, J. A., Gridley, G., Weiderpass, E., Nyrén, O. & Linet, M. Venous thromboembolism and cancer. *Lancet* **351**, 1077–1080 (1998).
3. Rickles, F. R. & Levine, M. N. Epidemiology of thrombosis in cancer. *Acta Haematol.* **106**, 6–12 (2001).
4. Lee, A. Y. Y. & Levine, M. N. Venous thromboembolism and cancer: risks and outcomes. *Circulation* **107** (suppl.), 117–121 (2003).
5. Vigna, E. & Naldini, L. Lentiviral vectors: excellent tools for experimental gene transfer and promising candidates for gene therapy. *J. Gene Med.* **2**, 308–316 (2000).
6. Trusolino, L. & Comoglio, P. M. Scatter-factor and semaphorin receptors: cell signalling for invasive growth. *Nature Rev. Cancer* **2**, 289–300 (2002).
7. Schmidt, L. et al. Germline and somatic mutations in the tyrosine kinase domain of the *MET* proto-oncogene in papillary renal carcinomas. *Nature Genet.* **16**, 68–73 (1997).
8. Di Renzo, M. F. et al. Expression of the *Met/HGF* receptor in normal and neoplastic human tissues. *Oncogene* **6**, 1997–2003 (1991).
9. Pennacchietti, S. et al. Hypoxia promotes invasive growth by transcriptional activation of the *met* protooncogene. *Cancer Cell* **3**, 347–361 (2003).
10. Park, M. et al. Mechanism of *met* oncogene activation. *Cell* **45**, 895–904 (1986).
11. Dull, T. et al. A third-generation lentivirus vector with a conditional packaging system. *J. Virol.* **72**, 8463–8471 (1998).
12. Follenzi, A., Sabatino, G., Lombardo, A., Boccaccio, C. & Naldini, L. Efficient gene delivery and targeted expression to hepatocytes *in vivo* by improved lentiviral vectors. *Hum. Gene Ther.* **13**, 243–260 (2002).
13. Ponder, K. P. et al. Mouse hepatocytes migrate to liver parenchyma and function indefinitely after intrasplenic transplantation. *Proc. Natl Acad. Sci. USA* **88**, 1217–1221 (1991).
14. Pfeifer, A. et al. Transduction of liver cells by lentiviral vectors: analysis in living animals by fluorescence imaging. *Mol. Ther.* **3**, 319–322 (2001).
15. Rickles, F. R., Levine, M. N. & Dvorak, H. B. in *Hemostasis and Thrombosis* (eds Colman, R. W., Hirsh, J., Marder, V. J., Clowes, A. & George, J. N.) 1132–1152 (Lippincott Williams & Wilkins, Philadelphia, 2000).
16. Muller-Berghaus, G., ten Cate, H. & Levi, M. Disseminated intravascular coagulation: clinical spectrum and established as well as new diagnostic approaches. *Thromb. Haemost.* **82**, 706–712 (1999).
17. Collen, D. The plasminogen (fibrinolytic) system. *Thromb. Haemost.* **82**, 259–270 (1999).
18. Smith, W. L., De Witt, D. L. & Garavito, R. M. Cyclooxygenases: structural, cellular, and molecular biology. *Annu. Rev. Biochem.* **69**, 145–182 (2000).
19. Friederich, P. W. et al. Novel low-molecular-weight inhibitor of PAI-1 (XR5118) promotes endogenous fibrinolysis and reduces postthrombolysis thrombus growth in rabbits. *Circulation* **96**, 916–921 (1997).
20. Gupta, R. A. & DuBois, R. N. Colorectal cancer prevention and treatment by inhibition of cyclooxygenase-2. *Nature Rev. Cancer* **1**, 13–20 (2001).
21. Dvorak, H. F. Tumors: wounds that do not heal. *N. Engl. J. Med.* **315**, 1650–1659 (1986).
22. Zeng, Q., McCauley, L. K. & Wang, C. Y. Hepatocyte growth factor inhibits anoikis by induction of activator protein 1-dependent cyclooxygenase-2. Implication in head and neck squamous cell carcinoma progression. *J. Biol. Chem.* **277**, 50137–50142 (2002).
23. FitzGerald, G. A. COX-2 and beyond: Approaches to prostaglandin inhibition in human disease. *Nature Rev. Drug Discov.* **2**, 879–890 (2003).
24. Sidenius, N. & Blasi, F. The urokinase plasminogen activator system in cancer: recent advances and implication for prognosis and therapy. *Cancer Metastasis Rev.* **22**, 205–222 (2003).
25. Nakamura, T. et al. Molecular cloning and expression of human hepatocyte growth factor. *Nature* **342**, 440–443 (1989).
26. Trusolino, L., Pugliese, L. & Comoglio, P. M. Interactions between scatter factors and their receptors: hints for therapeutic applications. *FASEB J.* **12**, 1267–1280 (1998).
27. Follenzi, A., Ailles, L. E., Bakovic, S., Geuna, M. & Naldini, L. Gene transfer by lentiviral vectors is limited by nuclear translocation and rescued by HIV-1 pol sequences. *Nature Genet.* **25**, 217–222 (2000).

28. Medico, E. *et al.* The tyrosine kinase receptors Ron and Sea control "scattering" and morphogenesis of liver progenitor cells *in vitro*. *Mol. Biol. Cell* **7**, 495–504 (1996).
29. Li, C. & Wong, W. H. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA* **98**, 31–36 (2001).
30. Oshima, M. *et al.* Chemoprevention of intestinal polyposis in the *Apcdelta716* mouse by rofecoxib, a specific cyclooxygenase-2 inhibitor. *Cancer Res.* **61**, 1733–1740 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M. Risio, E. David and their staffs for pathology, M. Cilli for mouse surgery, R. Albano, A. Ferraro and R. Lo Noce for technical assistance, and M. Belluardo for computational analysis. We thank A. Cignetto for secretarial assistance. G.S. is the recipient of an AIRC fellowship. This research was supported by AIRC, CNR-MIUR, FIRB-MIUR, MIUR-PRIN and the Foundations CRT and 'Compagnia di San Paolo'.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.B. (carla.boccaccio@irci.it).

X-inactivation profile reveals extensive variability in X-linked gene expression in females

Laura Carrel¹ & Huntington F. Willard²

¹Department of Biochemistry and Molecular Biology, Pennsylvania State University College of Medicine, Hershey, Pennsylvania 17033, USA

²Institute for Genome Sciences & Policy, Duke University, Durham, North Carolina 27708, USA

In female mammals, most genes on one X chromosome are silenced as a result of X-chromosome inactivation^{1,2}. However, some genes escape X-inactivation and are expressed from both the active and inactive X chromosome. Such genes are potential contributors to sexually dimorphic traits, to phenotypic variability among females heterozygous for X-linked conditions, and to clinical abnormalities in patients with abnormal X chromosomes³. Here, we present a comprehensive X-inactivation profile of the human X chromosome, representing an estimated 95% of assayable genes in fibroblast-based test systems^{4,5}. In total, about 15% of X-linked genes escape inactivation to some degree, and the proportion of genes escaping inactivation differs dramatically between different regions of the X chromosome, reflecting the evolutionary history of the sex chromosomes. An additional 10% of X-linked genes show variable patterns of inactivation and are expressed to different extents from some inactive X chromosomes. This suggests a remarkable and previously unsuspected degree of expression heterogeneity among females.

Establishment of X-inactivation requires a key *cis*-acting master locus that includes the non-coding *XIST* gene^{6,7}. How *XIST* RNA and other epigenetic modifications^{8–10} are directed to sites along the inactive X chromosome (X_i), and how inactivation spreads in *cis* over the 155 Mb X chromosome still remain poorly understood. That some X-linked genes should escape inactivation was first predicted for pseudoautosomal genes on the X and Y chromosomes that show equivalent dosage, with two active copies in both males and females¹¹. Because genes that escape inactivation lack at least some of the epigenetic alterations that characterize the rest of the chromosome¹², their identification and analysis are important for understanding the mechanics of the X-inactivation process.

We have previously developed an X-inactivation assay system to detect gene expression directly from the X_i in human female cells by determining relative expression levels of polymorphic alleles in

heterozygous fibroblasts that demonstrate non-random inactivation⁵. For this study, we developed a quantitative assay based on fluorescent, single-nucleotide primer-extension (Q-SNaPshot) to distinguish between alleles of expressed single-nucleotide polymorphisms (SNPs) along the X chromosome (see Methods). For each heterozygous human fibroblast sample, complementary DNA was amplified, and the relative expression of each allele was determined after normalization to genomic DNA amplification products (Fig. 1). As all of the samples used show complete non-random inactivation⁵, genes that show biallelic expression must escape inactivation, with the normalized allelic ratios indicating the level of X_i expression (relative to that on the active X chromosome, X_a), for example, *HDHD1A* and *RIBC1* in Fig. 1. In contrast, monoallelic expression indicates that a gene is subject to inactivation (for example, the *TMLHE* gene in Fig. 1). Ninety-four genes spanning the X chromosome were tested in a total of 40 human samples (Supplementary Table 1). Notably, only about 65% of the genes were subject to inactivation in all heterozygous samples (Fig. 2); a further 20% were inactivated in some, but not all samples, and 15% escaped inactivation in all samples. Surprisingly, among

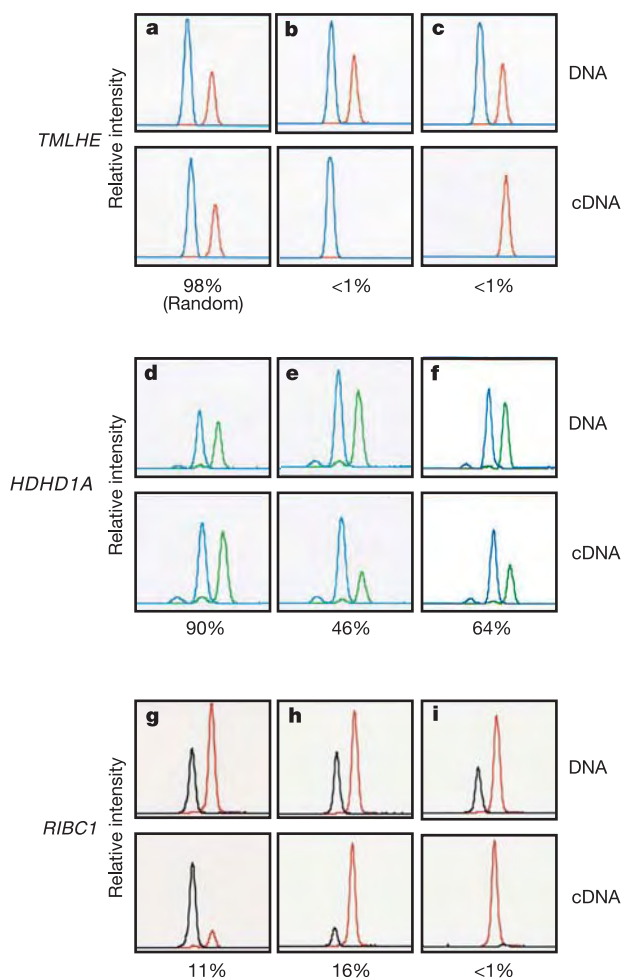


Figure 1 Q-SNaPshot assay of allelic expression for three X-linked genes. The assay measures expression by incorporating fluorescent dideoxy-nucleotides at polymorphic sites: G (blue), A (green), T (red), C (black). For each gene, DNA and cDNA amplification products are shown for three different cell lines (denoted a–i). For *TMLHE*, a randomly inactivated normal fibroblast line is included, whereas all other samples are non-randomly inactivated primary lines (that is, the same X chromosome is the X_i in all cells). Relative expression of the less intense allele for each sample (normalized with DNA to compensate for biased fluorescence output) is indicated as a percentage of the more intense allele, and by inference reflects the X_i expression level.

those genes that escape inactivation, most were not fully expressed from the X_i , demonstrating that escape from inactivation, even for pseudoautosomal genes, is generally partial and incomplete. For example, although the *FLJ39679* gene showed substantial X_i expression ($>75\%$ of X_a expression) in some cell lines, other cell lines showed X_i expression levels as low as $\sim 25\%$ of X_a expression levels (Fig. 2). These data indicate that females have considerable heterogeneity in levels of X-linked gene expression. At this point, however, we cannot establish whether this allelic variation is a specific feature of X-inactivation or reflects genome-wide variation.

To complement these data on a more comprehensive scale, we have established an X-inactivation profile in parallel with the complete sequence of the human X chromosome¹³. We determined the X-inactivation status of essentially all X-linked transcripts that are expressed in fibroblast-based assay systems^{4,5} (see Methods and Supplementary Table 2), and conclude that the current survey of 624 transcripts represents $\sim 95\%$ of assayable genes on the X chromosome.

X-inactivation status was assessed by analysing gene expression from human X_a and X_i chromosomes in an extensive panel of rodent/human somatic cell hybrids⁴. This method does not require distinction between X_a and X_i transcripts, and is applicable to all X-linked genes expressed in these cells. Each transcript was assayed

using polymerase chain reaction (PCR) with reverse transcription (RT-PCR) to determine whether expression is detected only from hybrid lines retaining X_a chromosomes (and is therefore a gene subject to X-inactivation) or from both X_a hybrid lines and X_i hybrid lines, indicating a gene that escapes X-inactivation. Genes were assayed in up to nine rodent/human somatic cell hybrids, each containing a different, normal human X_i chromosome (see Methods). Results for each gene in each X_i hybrid tested are depicted in Fig. 3a and are summarized for each transcript in Supplementary Table 3.

A total of 401 X-linked transcripts gave completely concordant results in all hybrids; that is, they were either expressed (74 transcripts) or silenced (327 transcripts) in all X_i hybrids tested. Considering only these transcripts (and excluding pseudoautosomal genes), the frequency of escape from inactivation for X-specific transcripts is minimally 16%, essentially identical to

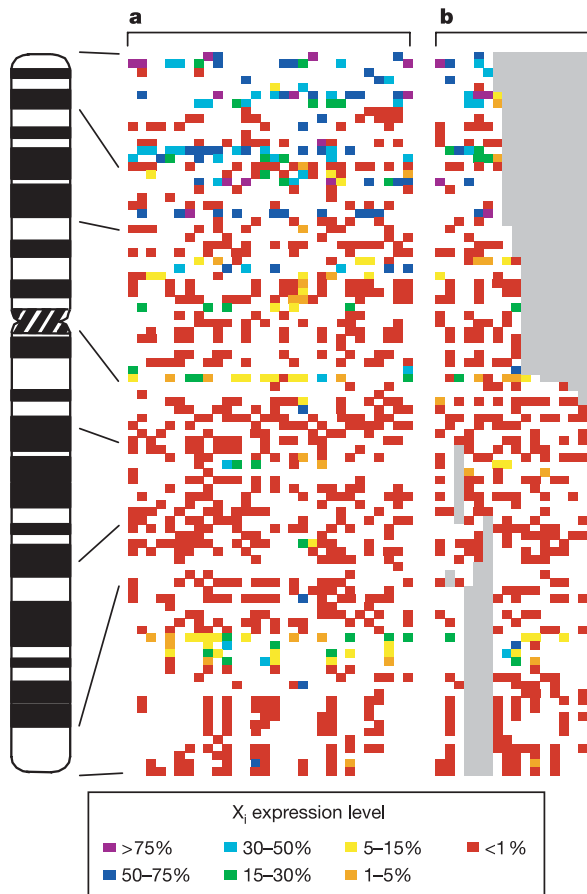


Figure 2 X-inactivation as a measure of allelic expression in non-randomly inactivated primary fibroblasts. **a**, **b**, Gene expression levels from normal X_i chromosomes (**a**) and from structurally abnormal X_i chromosomes (**b**). Each gene is linearly arrayed and approximate correspondence to chromosome location is indicated. The gene order is as in Supplementary Table 1. Each gene was assayed in all heterozygous individuals. Uninformative samples remain uncoloured. Colour-coding reflects relative X_i expression level as indicated. Grey boxes in **b** indicate absent portions of the X chromosome due to deletions or translocations.

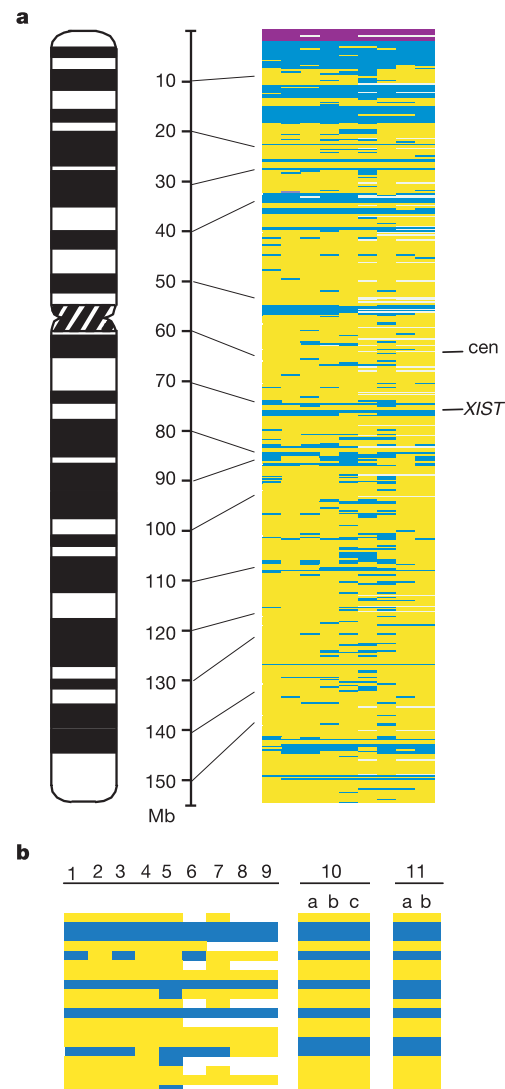


Figure 3 X-inactivation profile of the human X chromosome. **a**, 624 genes were tested in nine X_i hybrids. Each gene is linearly displayed. Blue denotes significant X_i gene expression, yellow shows silenced genes, pseudoautosomal genes are purple, and untested hybrids remain white. Positions of the centromere (cen) and *XIST* are indicated. **b**, To determine whether heterogeneity is largely a property of a specific chromosome, three independently derived hybrids from one X_i chromosome (denoted 10 a, b, c) and two hybrids carrying another X_i chromosome (11 a, b) were isolated, and results for 19 genes are shown adjacent to results for the original nine X_i chromosomes.

the estimate derived from studies in human cells. An additional 223 genes showed heterogeneity among different X_i hybrids and were expressed in some, but not all, of the X_i hybrids assayed (Fig. 3a), confirming and extending our observations in human cells (Fig. 2). Although this assay is inherently less quantitative than the SNP-based assay (see Methods), the majority of the genes escaping inactivation showed robust X_i expression; a minority were expressed from the X_i chromosome at relatively low levels (~ 10 – 15% that of the X_a). To assess the chromosomal basis for X_i heterogeneity, we analysed the same X chromosome isolated independently in multiple hybrids (Fig. 3b). Although the data set is limited, no variability was observed between different hybrids carrying the same X_i , supporting the idea that heterogeneity of X_i expression is largely an intrinsic property of the X chromosome tested, rather than an artefact of the hybrid model system. Of these heterogeneous genes, the majority (161 transcripts) gave concordant expression patterns in all but one or two hybrids, and for classification purposes⁴, were considered to escape (30 genes) or be subject to (131 genes) X-inactivation. In total, of the 612 X-specific transcripts classified in this manner, 458 (75%) are subject to inactivation in most or all of the X chromosomes tested, and 94 (15%) escape inactivation; the remaining 10% are heterogeneous (Fig. 3a and Supplementary Table 3). Combined, these last two categories of genes indicate that at least 25% of X-linked genes escape inactivation in all or a significant subset of X chromosomes tested, a result fully consistent with the data obtained with human cells (Fig. 2).

Importantly, given the limited number (<50) of X chromosomes screened in the two assays, the results from the primary human cell assay were in agreement with the X_i hybrid expression patterns in most instances. For example, 17 of 20 genes that escaped inactivation in X_i hybrids were expressed from the X_i in at least some of the human samples tested, although only 12 of these genes were expressed in all heterozygous lines tested (Supplementary Table 4). This comparison suggests that the total proportion of X-linked genes showing heterogeneous patterns of X_i expression in the general population may be even greater than the estimates provided here.

The distribution of genes that escape inactivation is clearly not random along the chromosome¹⁴. Large blocks of transcripts expressed from the X_i , apparent in both assays (Figs 2, 3), indicate that these genes are clustered and map primarily to the distal portion of the X chromosome short arm (Xp), perhaps related to their distance from the *XIST* gene, as recently proposed for the mouse¹⁵. Mammalian X and Y chromosomes are proposed to have diverged from an identical pair of ancestral autosomes¹⁶. Key to their evolution, structural rearrangements successively suppressed recombination, allowing independent evolution of each chromosome and leading to the largely degraded Y chromosome, which contains very few genes¹⁷. Evolution of the X chromosome, on the other hand, has been heavily influenced by X-chromosome inactivation, and such stringent regulation requires that gene content remain highly conserved¹⁶. As a consequence, the long arm (Xq) of the human X chromosome and a portion of proximal Xp, retain a high degree of synteny even with marsupial and monotreme X chromosomes¹⁸. By using the boundary of this X-conserved region and calculating sequence divergence between X and Y homologues, the locations of these ancestral rearrangements have been inferred¹⁹. The X-specific portion of the chromosome can be partitioned into five evolutionary strata that show increasing levels of sequence divergence with increasing distance from the distal tip of Xp^{13,19}. To explore an evolutionary influence on the X-inactivation profile, we compared X-inactivation patterns of genes mapping to each stratum (Fig. 4). Fulfilling the prediction that genes acquire the ability to be X-inactivated in response to the decay of Y-linked homologues²⁰, the proportion of genes that is expressed from the X_i in either the human samples (Fig. 4a) or in most hybrids (Fig. 4b) is

clearly highest in the youngest stratum (stratum 5) and lowest in the oldest stratum (stratum 1).

Even excluding pseudoautosomal genes, the majority of X-linked genes with Y homology escape X-inactivation and are expressed in all X_i hybrids tested (22 out of 35), and this proportion is even higher when considering X-linked genes with functional Y orthologues. Notably, however, two such genes—*TSPYL2* and *RBMX*—were subject to inactivation in all hybrids; these genes belong to the previously defined ampliconic class of XY-homologous genes¹⁷, suggesting that they may have less stringent dosage controls. We further examined the distribution and frequency of genes with Y homology, specifically among genes that were fully subject to or that fully escaped from inactivation (Supplementary Fig. 1). This comparison confirmed that younger strata contain a significantly higher proportion of genes with Y homology, but moreover, that the propensity of a gene with Y homology to escape inactivation diminishes in older strata, consistent with the idea that these genes lose the ability to escape inactivation over time²⁰.

The clustering of genes that escape inactivation, particularly evident in strata 4 and 5, supports the idea that control is at the level of chromosome domains, and is consistent with the premise that many genes escape inactivation because they lie within an epigenetic domain containing at least one X-linked gene with a Y homologue. However, this possibility cannot explain all genes that escape inactivation, because there are several other transcripts (particularly in stratum 1) that are expressed from the X_i even

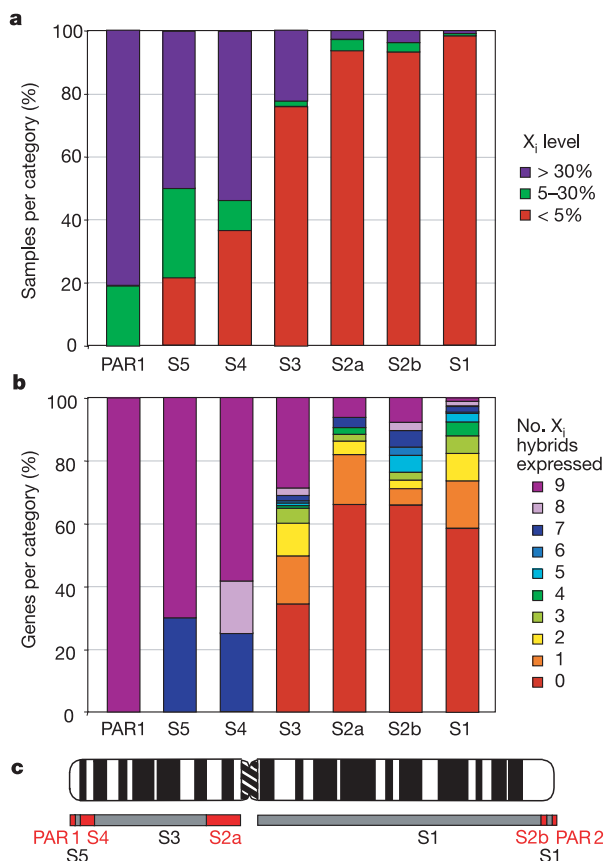


Figure 4 X_i expression data in primary fibroblasts and X_i hybrids correlate with location on X. Transcripts were subdivided according to X_i results and chromosomal region. **a**, The proportion of samples assayed in primary fibroblasts that show similar X_i results are indicated for the pseudoautosomal region PAR1, and for each X-specific evolutionary strata (S1–S5). **b**, For each chromosomal region, the percentage of genes showing similar X_i hybrid results (out of the nine X_i hybrid lines tested) are also shown. **c**, Location of each evolutionary strata on the X chromosome^{13,19,21}.

though they lack Y homology and map quite far from other expressed transcripts. One intriguing cluster of X_i -expressed genes maps within a gene-rich region of Xq28 that, on the basis of comparison with the chicken genomic sequence, has different evolutionary origins than the rest of Xq²¹. Using comparative mapping, this region is proposed to be included in stratum 2 (designated stratum 2b)²¹, potentially explaining the number of genes in this region of Xq28 that are expressed from all or at least a high percentage of the X_i chromosomes tested (Figs 2, 3).

Models to explain how inactivation spreads across the X chromosome postulate that specific sequences along the chromosome function as *cis*-acting 'booster elements'²². Lyon has proposed that long interspersed repeat elements (L1) could serve such a function²³, in part because of their overall twofold enrichment on the X chromosome relative to autosomes^{13,24}. L1 concentration is inversely correlated with the proportion of genes that escape inactivation¹³ (Supplementary Fig. 2). Stratum 5 has the highest proportion of genes that escape inactivation, but also the lowest L1 and the highest *Alu* repeat concentrations. In fact, L1 levels in the pseudoautosomal region and in strata 4 and 5 are below the genome average¹³. Locally, however, differences in L1 concentration that correlate with X-inactivation patterns are not apparent (Supplementary Fig. 3). At the very least, these data indicate that the genomic and epigenomic determinants are likely to be more complex than repeat content alone.

Another local factor that could influence inactivation status is the presence of a CpG island, as DNA methylation is well known to be involved in the maintenance of X_i silencing^{2,25}. In contrast to a previous report using a small number of genes²⁶, the whole chromosome profile shows that the distribution of island-associated genes does not differ between genes that escaped X-inactivation in all X_i hybrids and those that were completely silenced (Supplementary Fig. 4). However, the proportion of island-associated genes that are subject to or escape from inactivation does differ significantly from the proportion of genes containing CpG islands that demonstrate heterogeneous patterns of inactivation ($P < 0.001$ and $P < 0.025$ respectively). These data suggest that the presence of an island by itself does not determine whether a gene is inactivated; however, it seems that the absence of an island may contribute to the variability of X_i expression among different X chromosomes, as seen both in human cells and hybrids.

The large number of genes that escape inactivation, and their non-random distribution on the chromosome, has implications for counseling individuals with X chromosome abnormalities, estimated at 1 in 650 live births³. The evolutionary and genomic patterns of genes that escape inactivation suggest that aneuploidy for Xp is more severe than aneuploidy for Xq⁴. The profiles presented here provide a database of candidate genes to explain the clinical findings in such cases, although it is important to emphasize that the current profile has been obtained in fibroblast cells and may or may not reflect the situation in other tissues relevant to specific disease phenotypes.

Owing to different levels and different subsets of genes escaping X-inactivation, females will be even more variable with respect to X-linked gene expression than previously recognized. Because of these heterogeneous genes and the ~15% of genes that escape inactivation, the female genome differs from the male genome in at least four ways. First, the Y chromosome endows the male with at least several dozen genes that are absent in the female¹⁷. Second, the incomplete nature of X-inactivation means that at least 15% of X-linked genes are expressed at characteristically higher (but often variable) levels in females than in males. Third, a minimum of an additional 10% of genes show heterogeneous X-inactivation and thus differ in expression levels among females, whereas all males express a single copy of such genes. And fourth, the long-recognized random nature of X-inactivation indicates that females, but not

males, are mosaics of two cell populations with respect to X-linked gene expression^{1,3}.

Notwithstanding the genomic and biological significance of these sex-specific differences, their general clinical and phenotypic implications remain unexplored. However, as many of the genes that escape from inactivation do not have Y-linked homologues, strict dosage compensation may not be necessary for all genes on the chromosome. Such characteristic genomic differences should be recognized as a factor for explaining sex-specific phenotypes both in complex disease as well as in normal, sexually dimorphic traits²⁷. □

Methods

X-linked genes included in this study

Out of 931 X-linked genes annotated in NCBI build 34.3 (Supplementary Table 2), 267 were excluded from analysis because (1) they mapped to multiple locations, (2) they were computationally-predicted proteins that lacked significant expressed-sequence-tag (EST) support, or (3) they were transcripts with restricted expression patterns; for example, 10% of genes on the X chromosome are cancer-testis antigen genes that are not expressed in fibroblasts¹³ (Supplementary Table 2 and Supplementary Note 1). Primers were designed for the remaining genes, and expression was initially tested in DNA and cDNA from human and mouse fibroblasts and mouse/human hybrid cell lines carrying X_a chromosomes. This preliminary screening identified 471 transcripts that were expressed in primary fibroblasts and could therefore be tested for X-inactivation. This frequency of expressed transcripts (71%) is consistent with microarray studies of fibroblasts²⁸. An additional 153 transcripts (including full-length mRNAs and ESTs) that remain unassociated with currently annotated genes were also included in the profile. The transcripts analysed are listed in Supplementary Table 3.

Allelic expression in primary fibroblast cell lines

A panel of 40 primary fibroblast cell lines from females with non-random inactivation has been previously characterized³. These lines show complete skewing of inactivation, primarily as a result of secondary cell selection that maintains dosage despite structural rearrangements involving the X chromosome³. Confirmation that these lines were non-randomly inactivated was established by multiple assays³. Cell lines were maintained and DNA and cDNA isolated as described³.

Several assays were used to determine relative allele frequencies⁵. A novel quantitative assay (Q-SNaPshot, a modification of the commercial SNaPshot assay (Applied Biosystems)) was established that uses primer extension to incorporate a single fluorescent dideoxy nucleotide at the polymorphic site and quantitates products on an ABI 3100 (L.C., unpublished). Samples were normalized by comparison to heterozygous DNA samples with a known 1:1 allele ratio (Fig. 1). Quantitative conditions were confirmed by correctly predicting allele ratios for known mixtures of homozygous samples (data not shown). All assays were performed in duplicate.

Expression of transcripts from active and inactive X hybrids

Somatic cell hybrids, cell culture conditions, and nucleic acid purification methods have previously been reported⁴. Initial studies established that many genes have very low levels of X_i expression (Supplementary Fig. 5), and we therefore adopted a protocol that would allow us to identify significant X_i expression levels in hybrids. Large-scale reverse transcription reactions were performed as described (but volumes were increased tenfold⁴). RNA samples were treated with DNaseI before reverse transcription, to eliminate any contaminating genomic DNA. Each large-scale cDNA synthesis panel was normalized by amplification of the pseudoautosomal gene *MIC2*, which is equally expressed from X_a and X_i (Supplementary Fig. 5 and Supplementary Note 2).

Each gene was tested at two concentrations of cDNA, corresponding to ~50 ng and 250 ng of total RNA (Supplementary Fig. 6). Cycles were minimized to ensure that, in particular, the dilute cDNA samples were amplified in the exponential range. Although not strictly quantitative, this protocol allowed us to estimate the relative abundance of X_a and X_i transcripts. Faint amplification products that disappeared in the diluted samples were not considered positive. Samples that showed heterogeneous expression among different X_i chromosomes, or discordant patterns between cDNA dilutions, were repeated to ensure that scored positives were significantly expressed. Because of the semiquantitative nature of the RT-PCR assay used, the X_i expression levels in samples scored as positive are estimated to be at least at 10–15% of X_a levels (Supplementary Figs 5, 6 and Supplementary Note 2).

To assess the reproducibility of results within the same gene, ~15% of genes were assayed independently and blindly using at least two primer pairs. Virtually all (97%) of the X_i hybrid samples scored were in complete agreement; those that did not agree usually had faint bands that were called positive in one case but not another. For three genes, discordant results from more than one hybrid suggest that the assayed transcripts are incorrectly linked in the Unigene database and belong to distinct transcripts.

A small number of genes included in this profile have been tested by other laboratories (for example, see ref. 29). Most were re-analysed in our more extensive set of hybrids. The data here include and extend data previously reported from our laboratory⁴.

Transcript analysis

Transcripts were mapped physically along the X chromosome using the May 2004 assembly from the UCSC genome browser (<http://genome.ucsc.edu>). CpG islands, as

annotated in the UCSC genome browser, were assigned to each gene if they lay within 2 kb upstream or downstream of the 5' end of the RefSeq transcript or Unigene cluster. ESTs and most single exon transcripts were excluded from this analysis (Supplementary Table 3).

Evolutionary strata were assigned as described^{13,19,21}, and specific gene designations are indicated in Supplementary Table 3 and Supplementary Note 3. *Alu* and L1 repeat information was as annotated in the UCSC genome browser.

Received 24 January; accepted 17 February 2005; doi:10.1038/nature03479.

1. Lyon, M. F. Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature* **190**, 372–373 (1961).
2. Plath, K., Mlynarczyk-Evans, S., Nusinow, D. A. & Panning, B. *Xist* RNA and the mechanism of X chromosome inactivation. *Annu. Rev. Genet.* **36**, 233–278 (2002).
3. Willard, H. F. in *The Metabolic and Molecular Bases of Inherited Disease* 8th edn (eds Scriver, C. R., Beaudet, A. L., Sly, W. S., Valle, D., Childs, B. & Vogelstein, B.) 1191–1221 (McGraw-Hill, New York, 2000).
4. Carrel, L., Cottle, A. A., Goglin, K. C. & Willard, H. F. A first-generation X-inactivation profile of the human X chromosome. *Proc. Natl Acad. Sci. USA* **96**, 14440–14444 (1999).
5. Carrel, L. & Willard, H. F. Heterogeneous gene expression from the inactive X chromosome: an X-linked gene that escapes X inactivation in some human cell lines but is inactivated in others. *Proc. Natl Acad. Sci. USA* **96**, 7364–7369 (1999).
6. Brown, C. J. *et al.* A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature* **349**, 38–44 (1991).
7. Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S. & Brockdorff, N. Requirement for *Xist* in X chromosome inactivation. *Nature* **379**, 131–137 (1996).
8. Rougeulle, C. *et al.* Differential histone H3 Lys-9 and Lys-27 methylation profiles on the X chromosome. *Mol. Cell. Biol.* **24**, 5475–5484 (2004).
9. Chadwick, B. P. & Willard, H. F. Barring gene expression after *XIST*: maintaining facultative heterochromatin on the inactive X. *Semin. Cell Dev. Biol.* **14**, 359–367 (2003).
10. Chadwick, B. P. & Willard, H. F. Multiple spatially distinct types of facultative heterochromatin on the human inactive X chromosome. *Proc. Natl Acad. Sci. USA* **101**, 17450–17455 (2004).
11. Lyon, M. F. Sex chromatin and gene action in the mammalian X-chromosome. *Am. J. Hum. Genet.* **14**, 135–148 (1962).
12. Brown, C. J. & Gready, J. M. A stain upon the silence: genes escaping X inactivation. *Trends Genet.* **19**, 432–438 (2003).
13. Ross, M. T. *et al.* The DNA sequence of the human X chromosome. *Nature* doi:10.1038/nature03440 (this issue).
14. Miller, A. P. & Willard, H. F. Chromosomal basis of X chromosome inactivation: identification of a multigene domain in Xp11.21-p11.22 that escapes X inactivation. *Proc. Natl Acad. Sci. USA* **95**, 8709–8714 (1998).
15. Huynh, K. D. & Lee, J. T. Inheritance of a pre-inactivated paternal X chromosome in early mouse embryos. *Nature* **426**, 857–862 (2003).
16. Ohno, S. *Sex Chromosomes and Sex-linked Genes* (Springer, Berlin, 1967).
17. Skalketsky, H. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825–837 (2003).
18. Graves, J. A. Mammals that break the rules: genetics of marsupials and monotremes. *Annu. Rev. Genet.* **30**, 233–260 (1996).
19. Lahn, B. T. & Page, D. C. Four evolutionary strata on the human X chromosome. *Science* **286**, 964–967 (1999).
20. Jegalian, K. & Page, D. C. A proposed path by which genes common to mammalian X and Y chromosomes evolve to become X inactivated. *Nature* **394**, 776–780 (1998).
21. Kohn, M., Kehrer-Sawatzki, H., Vogel, W., Graves, J. A. & Hameister, H. Wide genome comparisons reveal the origins of the human X chromosome. *Trends Genet.* **20**, 598–603 (2004).
22. Gartler, S. M. & Riggs, A. D. Mammalian X-chromosome inactivation. *Annu. Rev. Genet.* **17**, 155–190 (1983).
23. Lyon, M. F. X-chromosome inactivation: a repeat hypothesis. *Cytogenet. Cell Genet.* **80**, 133–137 (1998).
24. Bailey, J. A., Carrel, L., Chakravarti, A. & Eichler, E. E. Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. *Proc. Natl Acad. Sci. USA* **97**, 6634–6639 (2000).
25. Mohandas, T., Sparkes, R. S. & Shapiro, L. J. Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation. *Science* **211**, 393–396 (1981).
26. Ke, X. & Collins, A. CpG islands in human X-inactivation. *Ann. Hum. Genet.* **67**, 242–249 (2003).
27. Committee on Understanding the Biology of Sex and Gender Differences. In *Exploring the Biological Contributions to Human Health—Does Sex Matter?* (eds Wizemann, T. M. & Pardue, M.) (National Academies Press, Washington, DC, 2001).
28. Iyer, V. R. *et al.* The transcriptional program in the response of human fibroblasts to serum. *Science* **283**, 83–87 (1999).
29. Esposito, T. *et al.* Escape from X inactivation of two new genes associated with DXS6974E and DXS7020E. *Genomics* **43**, 183–190 (1997).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M. Ross for sharing information before publication and gratefully acknowledge technical assistance from G. Nickel, K. Trevarthen, J. Dunn, A. Cottle and M. Moon. This work was supported in part by a National Institutes of Health research grant to H.F.W. and L.C.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to L.C. (lcarrel@psu.edu).

The vacuolar Ca^{2+} -activated channel TPC1 regulates germination and stomatal movement

Edgar Peiter¹, Frans J. M. Maathuis¹, Lewis N. Mills², Heather Knight³, Jérôme Pelloux¹, Alistair M. Hetherington² & Dale Sanders¹

¹Biology Department, Area 9, University of York, PO Box 373, York YO10 5YW, UK

²Biology Department, Lancaster Environment Centre, University of Lancaster, Lancaster LA1 4YQ, UK

³Department of Plant Sciences, University of Oxford, South Parks Road, Oxford OX1 3RB, UK

Cytosolic free calcium ($[\text{Ca}^{2+}]_{\text{cyt}}$) is a ubiquitous signalling component in plant cells¹. Numerous stimuli trigger sustained or transient elevations of $[\text{Ca}^{2+}]_{\text{cyt}}$ that evoke downstream stimulus-specific responses. Generation of $[\text{Ca}^{2+}]_{\text{cyt}}$ signals is effected through stimulus-induced opening of Ca^{2+} -permeable ion channels that catalyse a flux of Ca^{2+} into the cytosol from extracellular or intracellular stores. Many classes of Ca^{2+} current have been characterized electrophysiologically in plant membranes². However, the identity of the ion channels that underlie these currents has until now remained obscure. Here we show that the *TPC1* ('two-pore channel 1') gene of *Arabidopsis thaliana* encodes a class of Ca^{2+} -dependent Ca^{2+} -release channel that is known from numerous electrophysiological studies as the slow vacuolar channel^{3–5}. Slow vacuolar channels are ubiquitous in plant vacuoles, where they form the dominant conductance at micromolar $[\text{Ca}^{2+}]_{\text{cyt}}$. We show that a *tpc1* knockout mutant lacks functional slow vacuolar channel activity and is defective in both abscisic acid-induced repression of germination and in the response of stomata to extracellular calcium. These studies unequivocally demonstrate a critical role of intracellular Ca^{2+} -release channels in the physiological processes of plants.

In *A. thaliana* the *TPC1* gene (AGI code At4g03560) is the only representative of the TPC gene family⁶ that is also present with one or two members in other dicotyledonous plants⁷, monocotyledonous plants^{8,9}, gymnosperms (*Pinus pinaster*, GenBank accession number BX252040), mosses (*Physcomitrella patens*, GenBank accession number BQ039582), and animals¹⁰. The corresponding protein comprises two fused Shaker-like units, each with six transmembrane spans, a number of basic residues in the fourth transmembrane domain of each Shaker-like unit, a cytosolic linker region with two EF hands—that is, helix-loop-helix domains—suggestive of a role in Ca^{2+} binding (Fig. 1a), and several putative phosphorylation sites. Expression of plant TPC1 proteins in a yeast mutant lacking endogenous channel activity results in complementation of two phenotypic properties of the mutant—mating-induced cell death and low basal Ca^{2+} influx—suggesting that TPC1 might form a Ca^{2+} -permeable channel^{6–9}.

To determine the intracellular location of the TPC1 protein in *planta*, we constructed a TPC1–green fluorescent protein (GFP) fusion protein and introduced it into *Arabidopsis* mesophyll protoplasts. Figure 1b and c show images of protoplasts expressing GFP alone. In intact protoplasts, fluorescence appears as a clear band surrounding most of the cell, but is also present in a cytosolic-rich region that surrounds the nucleus and chloroplasts. Fluorescence rapidly disperses after disruption of protoplasts expressing GFP alone, suggesting a cytosolic location for the GFP (Fig. 1c). Images of cells expressing the TPC1–GFP fusion protein are shown in Fig. 1d and e. The intact protoplast exhibits GFP fluorescence that is distinct from the plasma membrane and strongly suggestive of an association with intracellular membranes (Fig. 1d). Intact vacuoles

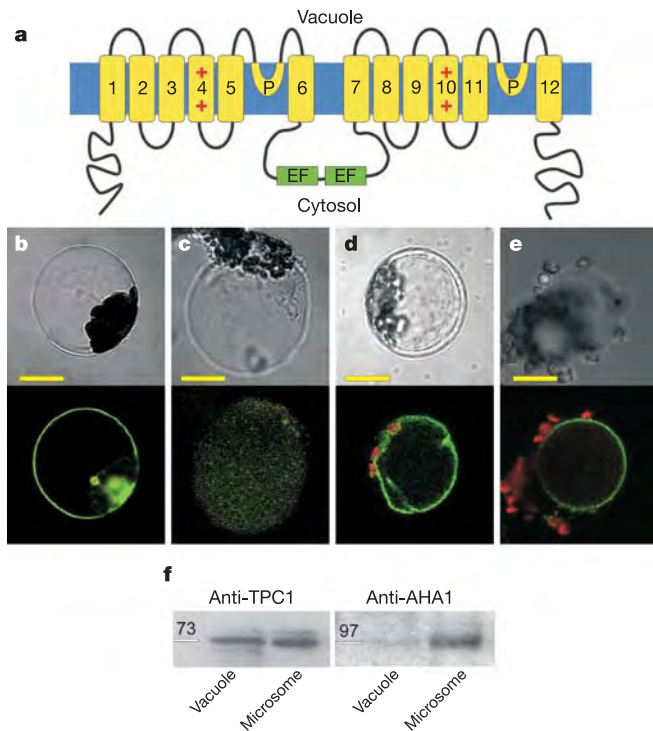


Figure 1 Topology and vacuolar localization of the *Arabidopsis* TPC1 protein. **a**, Topology of TPC1, showing two EF hands in the cytosolic linker (green), basic residues in the fourth transmembrane segment (TMS) of each domain (+), and putative pore domains (P) between the fifth and sixth TMS. **b–e**, Bright field (upper panels) and fluorescence images (lower panels) of protoplasts and vacuoles. GFP-specific fluorescence (green) was spectrally unmixed from autofluorescence (red). Scale bars: 20 μ m. **b**, GFP-transformed protoplast; **c**, imaged immediately after bursting. **d**, TPC1-GFP-transformed protoplast; **e**, vacuole isolated after bursting. **f**, Western blot using microsomal membranes and intact vacuoles, probed with antibodies against TPC1 and AHA2. Numbers indicate apparent molecular mass in kDa.

released from protoplasts show fluorescence that remains with the vacuolar membrane (Fig. 1e).

Further analyses confirmed the vacuolar location of TPC1: western blots comparing intact vacuole-derived membranes and microsomes prepared from *Arabidopsis* shoots clearly show the presence of TPC1 in the vacuolar fraction that is almost free of plasma membrane (Fig. 1f). In addition, proteomic analyses of the

vacuolar membrane from *Arabidopsis* suspension culture¹¹ or rosette leaves¹² identified TPC1 as the only annotated cation channel, a finding that we have independently verified with *Arabidopsis* shoot material (data not shown). The fact that vacuolar proteomics reveals TPC1 alone, and no other known vacuolar cation channels such as KCO1 (ref. 13), suggests that TPC1 is relatively abundant in plant vacuoles.

To identify which of the many previously reported vacuolar ion channel activities might be associated with the presence of TPC1, we investigated the electrophysiological properties of vacuoles from *Arabidopsis* plants varying in their level of TPC1 expression. For this purpose, we isolated a T-DNA knockout line, *tpc1-2*, from the Salk collection (Fig. 2a). This line has a truncated TPC1 transcript (Fig. 2b) and no detectable levels of wild-type TPC1 protein (Fig. 2c). TPC1 expression of another Salk T-DNA line, *tpc1-1*, was only knocked down by 70–80%, and TPC1 protein was detectable. Therefore, no further analyses were performed with *tpc1-1*. In addition, we created *Arabidopsis* plants constitutively overexpressing TPC1 under the control of the CaMV 35S promoter. The two TPC1-overexpressing lines used, 5.6 and 10.21, were found to have a 25 to 30-fold increase in TPC1 transcript levels when assessed by real-time polymerase chain reaction with reverse transcription (RT-PCR; Fig. 2d).

The presence of two Ca^{2+} -binding EF hands in the TPC1 protein, as well as its vacuolar location, prompted us to investigate whether TPC1 is, or forms part of, the slow vacuolar (SV) channel, which has a steep $[\text{Ca}^{2+}]_{\text{cyt}}$ dependence⁵. The SV channel is ubiquitous in plants and was one of the first plant ion channels to be characterized⁵. It forms the predominant vacuolar membrane conductance, and transports a wide range of cations, including calcium³. The SV channel is therefore believed to play a role in cation homeostasis and to function in Ca^{2+} -induced Ca^{2+} release from the vacuole^{4,14}. Channel activation is steeply voltage-dependent, typically over positive membrane voltages, although the activation mechanism is complicated and sensitive to many factors.

Figure 3a shows recordings of channel activity in an excised luminal-side-out membrane patch from a mesophyll vacuole of wild-type Col-0 *Arabidopsis*. Slowly activating, outward-rectifying currents were recorded only in the presence of high concentrations of Ca^{2+} ($>5 \mu\text{M}$) on the cytosolic side, as used here in the recording pipette. This steeply $[\text{Ca}^{2+}]_{\text{cyt}}$ -dependent (Fig. 3b) and voltage-dependent current is typical for plant vacuolar membranes and is diagnostic of SV channel activity⁵. In symmetrical, 100 mM KCl solutions, SV-type currents were observed in 17 of the 49 vacuoles patched (35%), and the unitary channel conductance was $54 \pm 6 \text{ pS}$ ($n > 250$ vacuoles). Figure 3c shows that under bi-ionic conditions (50 mM CaCl_2 and 100 mM KCl on the cytoplasmic and luminal

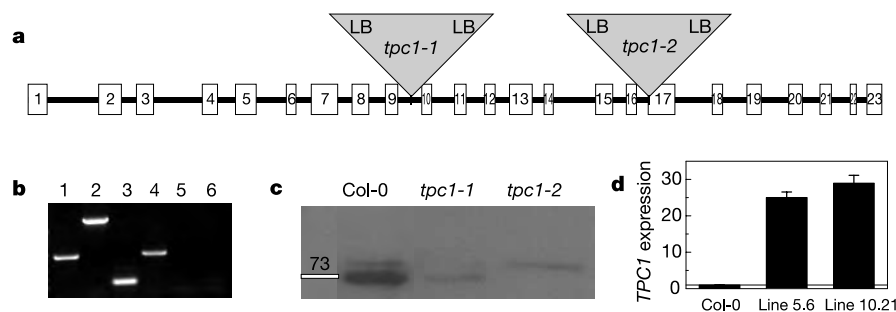


Figure 2 Characterization of TPC1-misexpressing *Arabidopsis thaliana* lines. **a**, Each of the inserts in *tpc1-1* and *tpc1-2*, located in the ninth intron and seventeenth exon respectively, is a concatameric T-DNA. **b**, The *tpc1-2* mutant lacks full-length TPC1 transcript. RT-PCR reactions with RNA isolated from Col-0 (lanes 1–3) and *tpc1-2* plants (4–6), using primers upstream (1, 4), across (2, 5), or downstream of the insertion site

(3, 6). **c**, As compared with Col-0 wild type, the level of TPC1 protein is reduced in *tpc1-1* and undetectable in *tpc1-2*. Number indicates apparent molecular mass in kDa. **d**, TPC1 expression levels in mature leaves of 35S::TPC1-transformed plants were increased 25- to 30-fold. Data are the means $\pm$ s.e.m. of four independent determinations from separate plants.

side of the vacuolar membrane, respectively) the single-channel conductance decreased to around 11 pS. In common with previous work¹⁵, this shows that in these conditions the Ca^{2+} flux is approximately eight times smaller than the K^{+} flux at comparable concentrations. However, on the basis of reversal potentials, relative $\text{Ca}^{2+}/\text{K}^{+}$ permeability ratios greater than one have been calculated^{3,4}. This disparity of permeability ratios is indicative of a multi-ion pore, a notion confirmed by the observation that the presence of both Ca^{2+} and K^{+} leads to anomalous mole fraction effects¹⁶.

Using identical recording conditions to those for Fig. 3a, Fig. 3d shows currents from a luminal side-out patch of a mesophyll vacuole from a *tpc1-2* mutant that was grown in identical conditions to the wild type. Clearly, there are no detectable time-dependent currents in the mutant lacking TPC1 protein. In total, recordings were obtained with 62 vacuoles from 10 different *tpc1-2* plants at various developmental stages. Time-dependent currents were not observed in any of these, although occasionally a small cation conductance (10–12 pS) channel with the hallmark characteristics of the vacuolar K^{+} (VK) channel⁴ was observed. The difference between the number of observed SV-type currents in wild-type and mutant plants is highly significant ($P < 10^{-7}$). Interestingly, previous work has demonstrated a decrease in SV-type currents in

a KCO1 channel mutant¹³, suggesting that KCO1 may be involved in the regulation of this conductance. In contrast, the complete abolition of SV currents in the *tpc1-2* mutant strongly suggests that TPC1 constitutes an essential component of the SV channel.

To confirm that the SV currents did indeed result from the presence of TPC1, we transformed *tpc1-2* mutant protoplasts with the *TPC1-GFP* construct and assayed transformed vacuoles for the presence of SV currents. Figure 3e clearly shows the presence of slowly activating outward-rectifying currents, demonstrating that the *TPC1-GFP* gene successfully complemented the mutant phenotype. In all, SV-type currents were observed in 11 of 27 vacuoles (41%) from *TPC1-GFP*-transformed protoplasts. Further demonstration of the role of the TPC1 protein in catalysing SV-type currents emerged from lines constitutively overexpressing *TPC1* under the control of the CaMV 35S promoter. Several lines were identified as having increased levels of *TPC1* transcript (Fig. 2d). Two of the lines were examined by patch clamp analysis, and in both the SV currents were markedly enhanced. In the example shown in Fig. 3f, currents in a luminal-side-out patch attain around 800 pA—considerably higher than commonly observed in the wild type (Fig. 3a). Figure 3g shows that the number of functional SV channels per patch was consistently higher in overexpressors: in

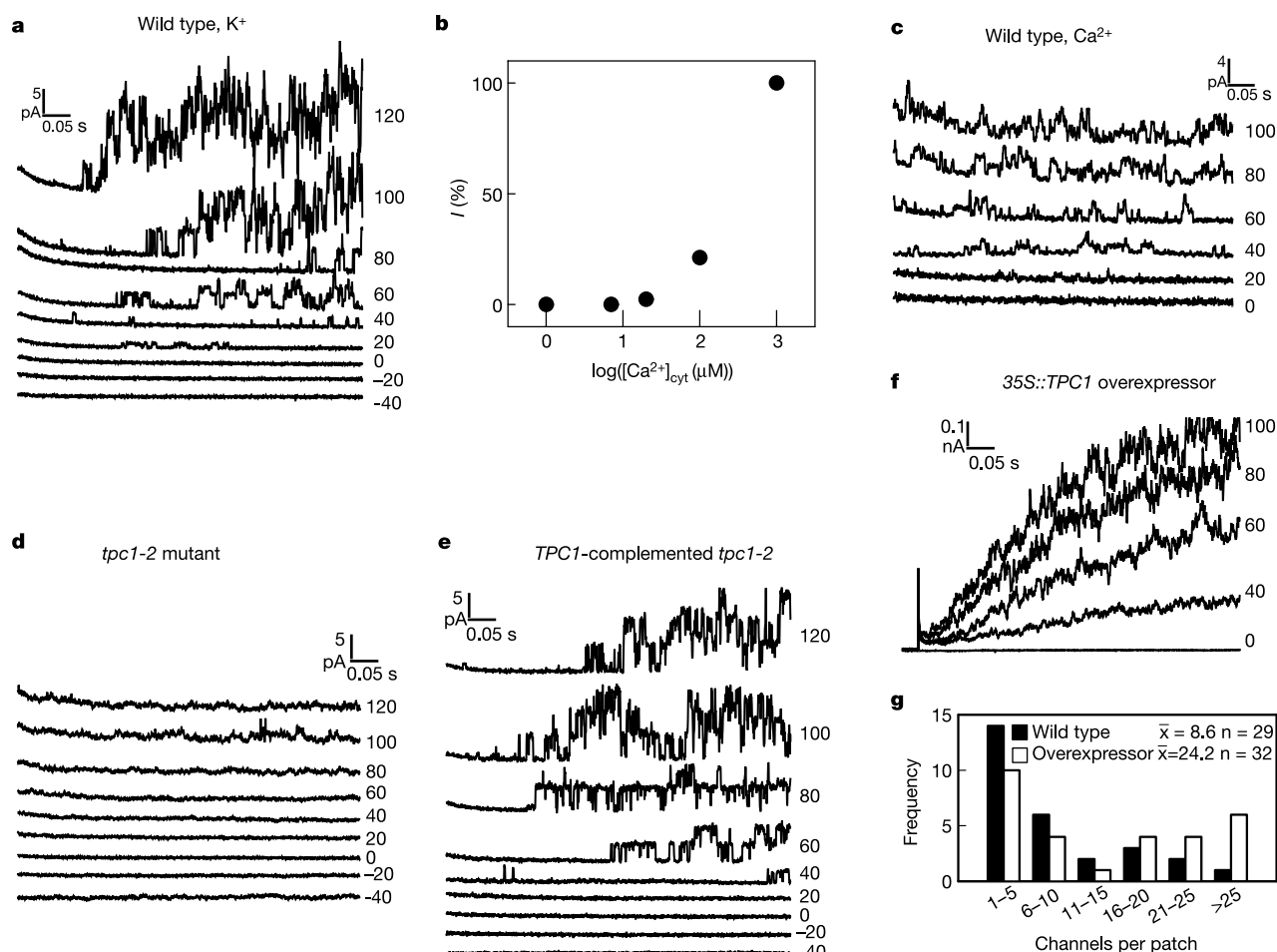


Figure 3 Electrophysiological characterisation of *Arabidopsis* mesophyll vacuolar membranes. **a**, Typical recording of slow vacuolar (SV) conductances in a luminal-side-out excised patch. Applied membrane potentials (in mV) are denoted on the right of these and succeeding traces. **b**, $[\text{Ca}^{2+}]_{\text{cyt}}$ -dependence of normalized current (I) at 100 mV. **c**, Typical recording under bi-ionic conditions (50 mM CaCl_2 and 100 mM KCl on cytoplasmic and luminal side, respectively). **d**, Typical recording from vacuoles derived

from the *tpc1-2* mutant showing complete absence of SV-type currents. **e**, Restoration of SV-type currents in vacuoles of the *tpc1-2* mutant transformed with *TPC1-GFP*. **f**, Increased numbers of SV-type channels in vacuoles isolated from constitutive *TPC1* overexpressors (line 5.6). **g**, Distribution of SV channel abundance per patch in vacuoles from Col-0 wild type and 35S::*TPC1* overexpressors.

wild-type vacuoles, around 9 channels per patch were recorded on average whereas in *TPC1*-overexpressing lines the average number of observed channels was around 24.

Previous studies have demonstrated that *TPC1* is ubiquitously expressed in all tissues of *Arabidopsis*: roots, flowers and leaves⁶, including guard cells¹⁷. To ascribe a physiological function to *TPC1*, we therefore analysed the lines lacking or overexpressing *TPC1* for a phenotype using conditions that are known to involve Ca^{2+} signalling in various tissues. The hormone abscisic acid (ABA) regulates many aspects of plant growth and development and has been shown in a number of studies to effect downstream responses via Ca^{2+} -dependent and Ca^{2+} -independent signalling networks^{18,19}. One process in which ABA plays a profound role is seed germination, where this stress hormone exerts a significantly inhibitory effect¹⁸. Figure 4a shows that wild type, mutant and overexpressors take a similar time to germinate in the absence of ABA. The *tpc1-2* mutant, however, yields a phenotype that is less sensitive to ABA, while two *TPC1*-overexpressing lines are ABA-hypersensitive over a wide range of ABA concentrations (Fig. 4a–c).

It has also been well established that regulation of stomatal aperture is Ca^{2+} -dependent, with closing stimuli such as ABA and extracellular Ca^{2+} evoking a rise in $[\text{Ca}^{2+}]_{\text{cyt}}$ that precedes the loss of turgor by the surrounding guard cells^{19,20}. Figure 4d shows that ABA applied over a range of concentrations inhibited opening to an identical extent in wild type and in the *tpc1-2* mutant. Similarly, no abnormal response was noted in the mutant with respect to ABA-promoted stomatal closure (data not shown). There are two possible explanations for these results. The first is that *TPC1*

is not involved in guard cell ABA signalling. Second, the robustness of guard cell ABA signalling, which exhibits properties reminiscent of scale-free networks²¹, might compensate for loss of *TPC1* through alternative cation-mobilizing systems. Figure 4e confirms the absence of a stomatal phenotype with respect to ABA in the *tpc1-2* mutant and in a *TPC1*-overexpressing line. But whereas wild-type and *TPC1*-overexpressing plants exhibited statistically significant ($P < 0.05$) reductions in aperture (relative to untreated epidermal peels) when treated with a high concentration of extracellular Ca^{2+} , the *tpc1-2* mutant was unresponsive. These observations show that while ABA is able to generate guard cell responses via *TPC1*-independent pathways, there is a requirement for *TPC1* in responding to elevated extracellular $[\text{Ca}^{2+}]$. In this respect, the *tpc1-2* guard cell phenotype mirrors that of a vacuolar H^+ -ATPase mutant, *det3* (ref. 22), which exhibits defective stomatal closure in response to external Ca^{2+} but not to ABA—a difference that is reflected in the respective responses of $[\text{Ca}^{2+}]_{\text{cyt}}$ to these closing stimuli.

The role of the SV channel in vacuolar Ca^{2+} release is still unresolved¹⁵ and, because the channel is relatively non-selective among mono- and divalent cations, the observed phenotypes might derive from *TPC1* functioning in general cation homeostasis and/or turgor regulation as well as in aspects of Ca^{2+} signalling. Indeed, it has been suggested that in guard cells SV channels mediate a large proportion of vacuolar K^+ release leading to closure of the pore^{2,3,16}. The balance between the dual functions—of cation homeostasis and Ca^{2+} release—will depend on the combination of the wide variety of factors known to alter the threshold activation voltage of the SV

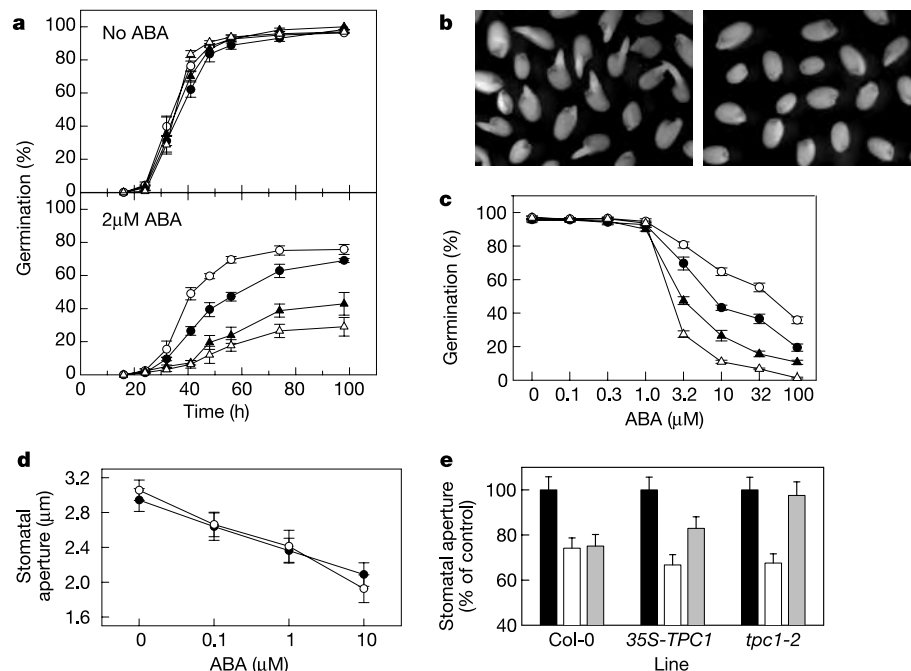


Figure 4 Phenotypic characterization of *tpc1-2* mutants and *TPC1* overexpressing lines. **a–c**, *TPC1* expression governs the ABA-sensitivity of seed germination. **a**, Time course of germination of Col-0 (filled circles), *tpc1-2* (open circles), and *35S::TPC1* (lines 5.6; filled triangles and 10.21; open triangles) seeds. Data are the means $\pm$ s.e.m. of a representative experiment (repeated three times with comparable results). **b**, Seeds of *tpc1-2* (left) and *35S::TPC1* line 5.6 (right) corresponding to the 48 h time points in the lower panel of **a**. **c**, Germination rate at different ABA concentrations, scored after four days' incubation. Symbols as in **a**. Data are the means $\pm$ s.e.m. of three independent experiments. **d**, *TPC1* expression is not required for inhibition of stomatal opening by ABA. Stomata of Col-0 (filled circles) and *tpc1-2* plants (open circles) were opened in buffer containing the indicated concentrations of ABA. Data are the means $\pm$ s.e.m. of a

representative experiment (40 stomatal aperture measurements per line and treatment). The experiment was repeated twice with comparable results. **e**, *TPC1* expression is required for Ca^{2+} -inhibition of stomatal opening. Stomata of Col-0, *tpc1-2* and *35S::TPC1* (line 5.6) plants were opened as above in buffer containing 100 μM CaCl_2 (black bars), 100 μM CaCl_2 and 10 μM ABA (white bars), or 10 mM CaCl_2 (grey bars). Data are the relative means $\pm$ s.e.m. of three independent experiments, each consisting of 60 stomatal aperture measurements per line and treatment. Absolute values of stomatal apertures in control medium (100 μM CaCl_2) were $3.51 \pm 0.15 \mu\text{m}$, $3.00 \pm 0.11 \mu\text{m}$, and $3.56 \pm 0.14 \mu\text{m}$ for Col-0, *tpc1-2*, and *35S::TPC1* plants, respectively.

channel. These factors include redox potential, phosphorylation state, Ca^{2+} /calmodulin, pH and free Mg^{2+} (ref. 2).

The present findings elucidate the molecular basis of a plant Ca^{2+} -permeable channel, and demonstrate roles played by this channel in essential physiological processes. The TPC1 channel is not only expressed ubiquitously in plants, but is also present in mammalian epithelia, where its function remains obscure¹⁰. The molecular identification of the plant SV channel now opens the way to address the detailed role of this ubiquitous class of cation channel in cell signalling. □

Methods

TPC1 knockout and overexpressor lines

Seven putative *A. thaliana* TPC1 T-DNA mutant lines were identified and retrieved from the Salk collection²³. PCR analysis and sequencing of the left border of the T-DNA insert confirmed the presence of an insert in two of these lines, *tpc1-1* (SALK_125658) and *tpc1-2* (SALK_145413). Real-time RT-PCR (ABI Prism 7000 sequence detection system, Applied Biosystems) using primers spanning the insertion site showed that transcript levels in *tpc1-1* were reduced by only 70–80%. In contrast, full-length transcripts were not detectable in *tpc1-2* (Fig. 2). For overexpression of TPC1, a TPC1 complementary DNA was cloned downstream of a CaMV 35S promoter into the pBI121 vector, from which the *GUS* coding sequence had been removed. *Arabidopsis* was transformed with the pBI121-TPC1 vector by floral dip²⁴. TPC1 overexpression levels were determined by real-time RT-PCR, using *UBQ10* as a constitutive control.

TPC1-GFP construct, protoplast transformation and GFP imaging

An *mGFP5* cDNA was amplified using primers containing the *Bam*HI site and inserted into the vector pART7 (ref. 25) to give the pART7GFP vector. A cDNA of TPC1 without a stop codon was amplified from *A. thaliana* Col-0 by RT-PCR using forward and reverse primers containing *Xho*I and *Xma*I sites, respectively, and inserted into the pART7GFP vector to give the pART7TPC1GFP vector encoding a carboxy-terminal fusion of TPC1 onto GFP under control of a CaMV 35S promoter. *Arabidopsis* mesophyll protoplasts were isolated and transformed with pART7GFP and pART7TPC1GFP plasmids as described²⁶. After 24–48 h of incubation, GFP fluorescence of protoplasts and vacuoles was observed by confocal laser scanning microscopy using a Zeiss LSM510 Meta based on an Axiovert 200M microscope (Zeiss). GFP fluorescence was determined by operating the microscope in lambda mode and spectrally unmixing the images using pre-recorded spectra of GFP and autofluorescence.

Germination assays

Surface-sterilized *Arabidopsis* seeds of identical age were cold-stratified (4 °C) in the dark for three days in de-ionized water containing 10 mM MES-KOH (pH 6.0) and ABA as indicated in Fig. 4a–c. The seeds were plated on agar containing the same concentrations of MES-KOH and ABA. Emergence of radicals was scored using a stereomicroscope. Experiments consisted of three replicate plates with 50 seeds per line and treatment, and were repeated three times.

Stomatal aperture bioassays

Epidermal strips of fully expanded leaves from 5-week-old *Arabidopsis* plants grown under short-day conditions were floated in the dark on CO_2 -free buffer containing 10 mM MES-KOH (pH 6.15). To promote the opening of stomata, epidermal strips were placed onto illuminated and aerated CO_2 -free buffer containing 50 mM KCl and 10 mM MES-KOH (pH 6.15), and 100 μM CaCl_2 unless indicated otherwise. Strips were incubated in white light for 2.5 h at 22 °C and were subsequently mounted on glass slides for determination of stomatal apertures using a Leica DM IRB microscope.

Electrophysiology

A. thaliana mesophyll protoplasts were isolated as described²⁷. Vacuoles were released from protoplasts by washing protoplasts with a solution containing 10 mM EDTA, 10 mM EGTA, pH 8 with an osmolality of 350 mOsm. General patch-clamp methodology was as described²⁸. Symmetrical solutions used in Fig. 3a, c–e contained 100 mM KCl, 2 mM CaCl_2 and 5 mM MES-Tris (pH 6.5). Experimental solutions were adjusted to 400 mOsm with sorbitol. The average number of channels recorded in Fig. 3f was calculated by dividing the total current at 100 mV by 5 pA (the approximate current through one 54 pS channel at 100 mV).

Membrane isolation and western blot analysis

Microsomal membranes of 6-week-old whole *Arabidopsis* shoots were isolated as described²⁹. To isolate vacuolar membranes, intact vacuoles were first collected from 6-week-old whole *Arabidopsis* shoots as described³⁰. Identical amounts of protein (2 μg per lane) were resolved by SDS–polyacrylamide gel electrophoresis (SDS–PAGE), transferred onto nitrocellulose membranes (Optitran BA-S 85, Schleicher & Schuell) by semidry-blotting, and probed with rabbit antibody (Eurogentec) raised against the amino-terminus of *Arabidopsis* TPC1 (MEDPLIGRDSLGGGG) or *Arabidopsis* AHA2 H^+ -ATPase (gift of M. Palmgren). The horseradish peroxidase-conjugated secondary antibody was detected by chemiluminescence (ECL, Amersham).

Received 3 November 2004; accepted 20 January 2005; doi:10.1038/nature03381.

- Sanders, D., Pelloux, J., Brownlee, C. & Harper, J. F. Calcium at the crossroads of signaling. *Plant Cell* **14**, S401–S417 (2002).
- White, P. J. Calcium channels in higher plants. *Biochim. Biophys. Acta* **1465**, 171–189 (2000).
- Allen, G. J. & Sanders, D. Vacuolar ion channels of higher plants. *Adv. Bot. Res.* **25**, 217–252 (1997).
- Ward, J. M. & Schroeder, J. I. Calcium-activated K^+ channels and calcium-induced calcium release by slow vacuolar ion channels in guard cell vacuoles implicated in the control of stomatal closure. *Plant Cell* **6**, 669–683 (1994).
- Hedrich, R. & Neher, E. Cytoplasmic calcium regulates voltage-dependent ion channels in plant vacuoles. *Nature* **329**, 833–836 (1987).
- Furuichi, T., Cunningham, K. W. & Muto, S. A putative two-pore channel AtTPC1 mediates Ca^{2+} flux in *Arabidopsis* leaf cells. *Plant Cell Physiol.* **42**, 900–905 (2001).
- Kadota, Y. *et al.* Identification of putative voltage-dependent Ca^{2+} -permeable channels involved in cryptogin-induced Ca^{2+} transients and defense responses in tobacco BY-2 cells. *Biochem. Biophys. Res. Commun.* **317**, 823–830 (2004).
- Hashimoto, K., Saito, M., Matsuoka, H., Iida, K. & Iida, H. Functional analysis of a rice putative voltage-dependent Ca^{2+} channel, OsTPC1, expressed in yeast cells lacking its homologous gene *CCH1*. *Plant Cell Physiol.* **45**, 496–500 (2004).
- Kurusu, T., Sakurai, Y., Miyao, A., Hirochika, H. & Kuchitsu, K. Identification of a putative voltage-gated Ca^{2+} -permeable channel (OsTPC1) involved in Ca^{2+} influx and regulation of growth and development in rice. *Plant Cell Physiol.* **45**, 693–702 (2004).
- Ishibashi, K., Suzuki, M. & Imai, M. Molecular cloning of a novel form (two-repeat) protein related to voltage-gated sodium and calcium channels. *Biochem. Biophys. Res. Commun.* **270**, 370–376 (2000).
- Szponarski, W., Sommerer, N., Boyer, J.-C., Rossignol, M. & Gibrat, R. Large-scale characterization of integral proteins from *Arabidopsis* vacuolar membrane by two-dimensional liquid chromatography. *Proteomics* **4**, 397–406 (2004).
- Carter, C. *et al.* The vegetative vacuole proteome of *Arabidopsis thaliana* reveals predicted and unexpected proteins. *Plant Cell* **16**, 3285–3303 (2004).
- Schönknecht, G. *et al.* KCO1 is a component of the slow-vacuolar (SV) ion channel. *FEBS Lett.* **511**, 28–32 (2002).
- Bewell, M. A., Maathuis, F. J. M., Allen, G. J. & Sanders, D. Calcium-induced calcium release mediated by a voltage-activated cation channel in vacuolar vesicles from red beet. *FEBS Lett.* **458**, 41–44 (1999).
- Pottosin, I. I., Dobrovinskaya, O. R. & Muniz, J. Conduction of monovalent and divalent cations in the slow vacuolar channel. *J. Membr. Biol.* **181**, 55–65 (2001).
- Allen, G. J. & Sanders, D. Control of ionic currents in guard cell vacuoles by cytosolic and luminal calcium. *Plant J.* **10**, 1055–1069 (1996).
- Leonhardt, N. *et al.* Microarray expression analyses of *Arabidopsis* guard cells and isolation of a recessive abscisic acid hypersensitive protein phosphatase 2C mutant. *Plant Cell* **16**, 596–615 (2004).
- Finkelstein, R. R., Gampala, S. S. L. & Rock, C. D. Abscisic acid signaling in seeds and seedlings. *Plant Cell* **14**, S15–S45 (2002).
- Schroeder, J. I., Allen, G. J., Hugouvieux, V., Kwak, J. M. & Waner, D. Guard cell signal transduction. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **52**, 627–658 (2001).
- McAinsh, M. R., Webb, A. A. R., Taylor, J. E. & Hetherington, A. M. Stimulus-induced oscillations in guard cell cytoplasmic free calcium. *Plant Cell* **7**, 1207–1219 (1995).
- Hetherington, A. M. & Woodward, F. I. The role of stomata in sensing and driving environmental change. *Nature* **424**, 901–908 (2003).
- Allen, G. J. *et al.* Alteration of stimulus-specific guard cell calcium oscillations and stomatal closing in *Arabidopsis det3* mutant. *Science* **289**, 2338–2342 (2000).
- Alonso, J. M. *et al.* Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science* **301**, 653–657 (2003).
- Clough, S. J. & Bent, A. F. Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J.* **16**, 735–743 (1998).
- Glave, A. P. A versatile binary vector system with a T-DNA organisational structure conducive to efficient integration of cloned DNA into the plant genome. *Plant Mol. Biol.* **20**, 1203–1207 (1992).
- Abel, S. & Theologis, A. Transient transformation of *Arabidopsis* leaf protoplasts: a versatile experimental system to study gene expression. *Plant J.* **5**, 421–427 (1994).
- Maathuis, F. J. M. & Sanders, D. Patch-clamping plant cells. *Methods Cell Biol.* **49**, 293–302 (1995).
- Maathuis, F. J. M. & Sanders, D. Sodium uptake in *Arabidopsis* roots is regulated by cyclic nucleotides. *Plant Physiol.* **127**, 1617–1625 (2001).
- Muir, S. R. & Sanders, D. Inositol 1,4,5-triphosphate-sensitive Ca^{2+} release across nonvacuolar membranes in cauliflower. *Plant Physiol.* **114**, 1511–1521 (1997).
- Höfte, H., Faye, L., Dickinson, C., Herman, E. M. & Chrispeels, M. J. The protein-body proteins phytohemagglutinin and tonoplast intrinsic protein are targeted to vacuoles in leaves of transgenic tobacco. *Planta* **184**, 431–437 (1991).

Acknowledgements We thank M. Fischer and R. V. Jolley for initial experiments, L. Skiera and S. Kilmartin for technical assistance, P. O'Toole for advice on confocal microscopy, G. Park for preparing the pART7GFP construct, M. G. Palmgren for providing AHA2 antibodies, and J. I. Schroeder for discussions. This work was funded by the Biotechnology and Biological Sciences Research Council and an award by the Koerber Foundation to D.S.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.S. (ds10@york.ac.uk). Sequence of *A. thaliana* Col-0 TPC1 is deposited in GenBank under accession number AF360372.

Molecular determinants and guided evolution of species-specific RNA editing

Robert A. Reenan

Department of Genetics and Developmental Biology, University of Connecticut Health Center, Farmington, Connecticut 06030, USA

Most RNA editing systems are mechanistically diverse, informationally restorative, and scattershot in eukaryotic lineages¹. In contrast, genetic recoding by adenosine-to-inosine RNA editing seems common in animals; usually, altering highly conserved or invariant coding positions in proteins^{2–4}. Here I report striking variation between species in the recoding of *synaptotagmin I* (*sytl*). Fruitflies, mosquitoes and butterflies possess shared and species-specific *sytl* editing sites, all within a single exon. Honeybees, beetles and roaches do not edit *sytl*. The editing machinery is usually directed to modify particular adenosines by information stored in intron-mediated RNA structures^{5–7}. Combining comparative genomics of 34 species with mutational analysis reveals that complex, multi-domain, pre-mRNA structures solely determine species-appropriate RNA editing. One of these is a previously unreported long-range pseudoknot. I show that small changes to intronic sequences, far removed from an editing site, can transfer the species specificity of editing between RNA substrates. Taken together, these data support a phylogeny of *sytl* gene editing spanning more than 250 million years of hexapod evolution. The results also provide models for the genesis of RNA editing sites through the stepwise addition of structural domains, or by short walks through sequence space from ancestral structures.

RNA editing systems programmatically alter messenger RNA sequences after transcription from genomic templates and are found enigmatically scattered among phyla. One example, base modification through the hydrolytic deamination of adenosine to inosine (A-to-I) by ADARs (for 'adenosine deaminases acting on RNAs'), can result in informational recoding: the ribosome interprets inosine as guanosine⁸. Curiously, this recoding occurs almost exclusively in gene products whose primary function is fast neuronal signalling³, in keeping with the observed neurological phenotypes of ADAR-deficient *Caenorhabditis elegans*, *Drosophila* and mice^{9–11}. Further, human neurological disease has been associated with altered gene recoding^{12,13}. The biological consequence of ADAR action at a particular site can vary markedly between genes and between species. Certain mammalian ionotropic glutamate receptor (iGluR) genes recode a functionally critical glutamine (Q) codon to that of arginine (R). Studies involving animals genetically altered in this recoding event are revealing. For instance, the GluR-2 (Q/R) site must be edited at nearly 100% or mice develop epilepsy and die postnatally¹⁴, whereas lack of editing at the GluR-6 (Q/R) site (normally about 75%) is relatively benign¹⁵, resulting in modest changes in synaptic plasticity and seizure vulnerability. *C. elegans* iGluR genes do not edit the same conserved amino acid; however, when an R-encoding version is introduced into worms, neurotoxicity and lethality ensue¹⁶.

Previous reports indicate that A-to-I RNA editing varies between arthropod species^{3,17,18}. Although the mechanism of gene recoding frequently involves imperfect base pairing of exonic and intronic sequences^{5–7,19}, the molecular basis of species-specific editing is unknown. *Drosophila synaptotagmin I* (*dsytI*), the Ca²⁺ sensor for synchronous neurotransmitter release²⁰, is a target of A-to-I RNA editing³. One exon, whose boundaries are conserved in all synaptotagmins²¹, was shown to possess four editing sites (A to D) that

recode highly conserved positions. To investigate whether editing varied between species, I performed reverse-transcriptase polymerase chain reaction (RT-PCR) to obtain *sytl* complementary DNAs from *Anopheles gambiae* (malaria mosquito), *Manduca sexta* (tobacco hawkmoth), *Apis mellifera* (honeybee), *Tribolium castaneum* (red flour beetle) and *Blattella germanica* (German cockroach). Direct sequence analysis of RT-PCR amplification products was performed to identify RNA editing sites, as described previously³. Editing was not detectable in *A. mellifera*, *T. castaneum* or *B. germanica sytl* genes. The remaining arthropods edited *sytl*, but no two species possessed the same set of editing sites (Fig. 1a, b). All

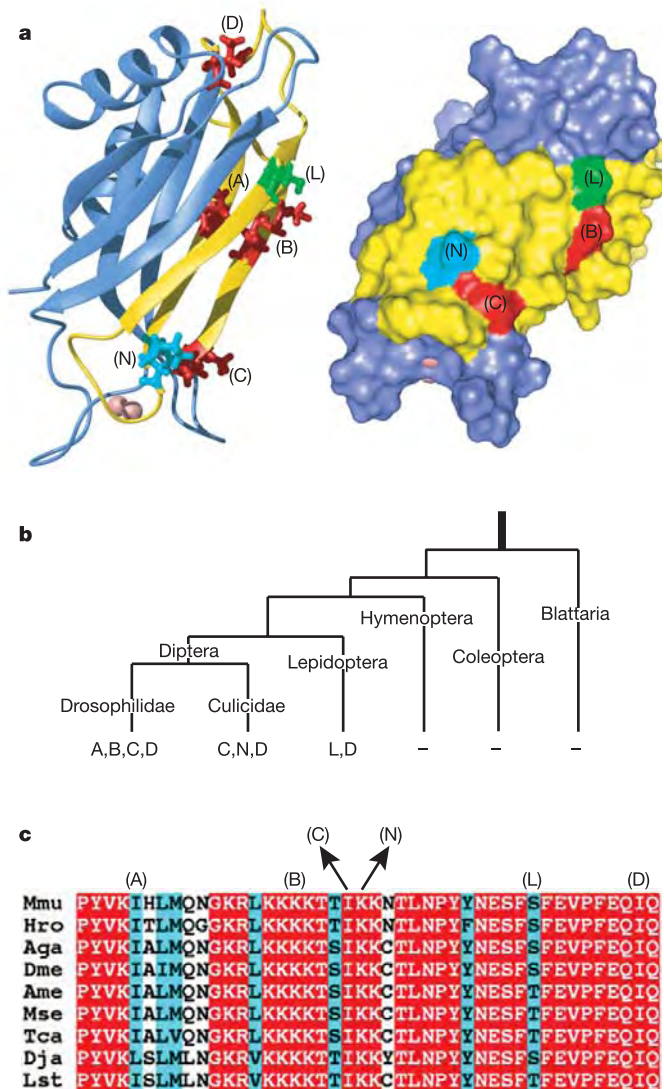


Figure 1 Species-specific genetic recoding of *synaptotagmin I*. **a**, Left: ribbon structure of the synaptotagmin I C₂B domain. Yellow corresponds to the conserved exon found in all *sytl* orthologues. Side chains of *D. melanogaster sytl* editing sites A–D (red), mosquito-specific site N (blue) and lepidopteran-specific site L (green) are indicated. Right: electrostatic surface potential showing surface-projecting residues of editing sites B, C, N and L (colours as in **a**). **b**, Cladogram of taxa in this study and editing sites present: Diptera (*Drosophilidae* (*Drosophila melanogaster*), *Culicidae* (*Anopheles gambiae*)), Lepidoptera (*Manduca sexta*), Hymenoptera (*Apis mellifera*), Coleoptera (*Tribolium castaneum*) and Blattaria (*Blattella germanica*). Dashes indicate no detectable editing. **c**, Protein sequence alignment of *synaptotagmin I* editing exon orthologues of the animal kingdom. Shaded invariant (red) and conserved (blue) amino acid residues are shown. Editing sites A–D, L and N result in the following recoding events (single-letter amino acid codes): site A, I → V; site B, K → R; site C, I → V; site D, I → M; site L, T → A; site N, K → R.

had a common editing site (D). In addition, *D. melanogaster* and *A. gambiae* shared a site (C). Finally, each species was found to possess species-specific editing sites (A, B, N and L). That species-specificity of RNA editing was a conserved feature among closely related species was confirmed by analysing *sytl* editing in additional representatives within each taxonomic group (Supplementary Table S1).

These editing sites in insects alter invariant or highly conserved residues within the SytI C2B domain (Fig. 1c), a Ca^{2+} -dependent phospholipid-binding machine essential for the rapid and synchronous release of neurotransmitter. Three species-specific editing sites (B, N and L) as well as shared editing site C recode amino acids positioned on an interaction surface that is crucial for proper SytI function^{22–25} (Fig. 1a). *Aplysia californica* (sea hare) employs alternative splicing of the same exon, generating functionally distinct SytI isoforms²⁶. A single amino acid difference accounted for differential function of these isoforms, corresponding to editing site L, and resulting in nearly the same amino acid change (Thr → Gly by alternative splicing, versus Thr → Ala by RNA editing).

Conservation of specific ADAR modification between related species can result in the simultaneous conservation of *cis* elements that direct RNA structure formation^{17,27–29}. To test this rule for *sytl*, genomic sequences were next cloned and sequenced for the genomic region spanning the editing sites from ten members of the family Drosophilidae with estimated divergence times ranging from 15 million to 80 million years ago. Sequence alignment revealed

two invariant intronic elements, E1 (33 nucleotides) and E2 (48 nucleotides) (Fig. 2a, Supplementary Fig. S1). The remaining intron sequences were highly divergent, although interelement spacing was a conserved feature.

To determine whether the E1/E2-containing intron directs the RNA editing of *Drosophila sytl*, the sequences spanning the edited exon, the intron and the downstream exon were expressed as a minigene in *Drosophila* S2 cells together with *Drosophila* ADAR (dADAR). Efficient and specific editing was observed with the use of a restriction-enzyme assay for sites C and D (see Methods), demonstrating the sufficiency of these sequences to direct ADAR modification (Fig. 2c, WT). No editing was observed at inappropriate adenosines or in the absence of co-transfected dADAR (data not shown). RNA structural computations with MFOLD consistently predicted two mutually exclusive lowest-energy structures: one pairing E1 with the upstream exonic region of editing sites B and C (domain I), and one pairing the downstream distal E2 with the exonic region of editing site D (domain II). The assumption of a pseudoknot structure reconciled these two duplexes in one structure (Fig. 2b). To probe this hypothetical structure, potentially disruptive mutations were introduced into domains I and II (Fig. 2b, c). Mutations M1 and M2 singly abolished editing at site D, whereas editing at site C occurred normally. Likewise, mutations M3 and M4 singly abolished editing at site C, whereas editing at site D was unaffected. Structurally compensatory double mutations M12 and M34 each restored editing in their respective domains, thus validating numerous predicted base-pair interactions. Because

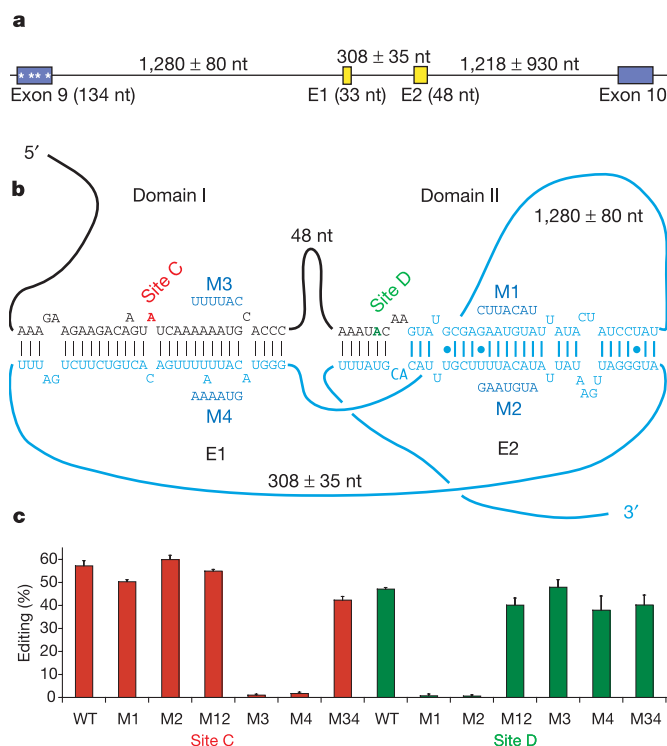


Figure 2 *Drosophila sytl* pre-mRNA forms a pseudoknot. **a**, Genomic organization of *Drosophilid sytl* gene editing sites. Exons (blue boxes), conserved elements E1 and E2 (yellow) and intron (line) shown with spacings as indicated (means $\pm$ s.d.). **b**, Predicted pseudoknot domain structure of *dsytI* pre-mRNA. Exon (black), intron (blue), and editing site C (red) and D (green) sequences are indicated. Mutations introduced into domains I and II are indicated above or below mutated sequences (M1–M4). **c**, Effects of mutations on the editing of sites C (red) and D (green) are indicated for disruptive single mutations (M1–M4) and compensatory double mutations (M12 and M34). WT, wild type. Values are means $\pm$ s.d.

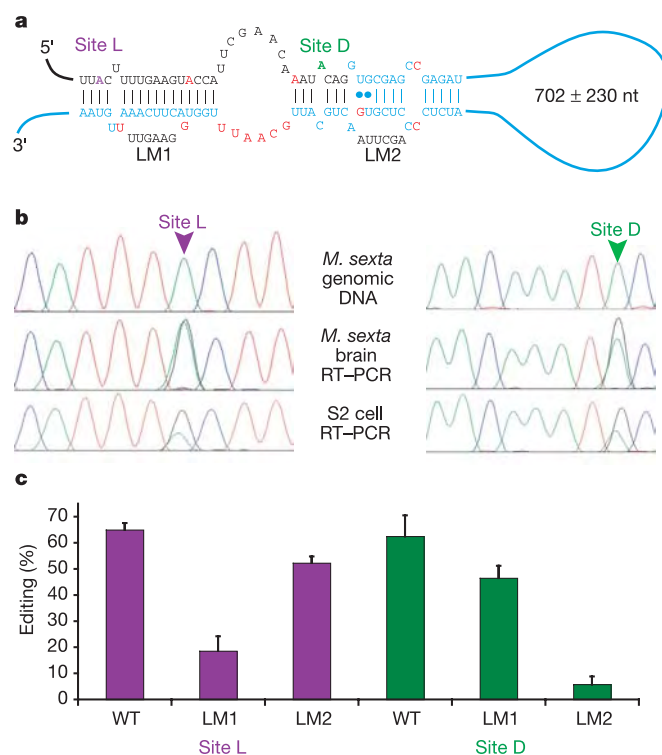


Figure 3 Structure of lepidopteran *sytl* editing site and heterologous editing by dADAR. **a**, Predicted structure of lepidopteran *sytl* pre-RNA. Exon (black), intron (blue) and variant positions (red) are indicated (see Supplementary Fig. S3). Variation in loop sequence length is indicated (mean $\pm$ s.d.). Intronic mutations are indicated (LM1 and LM2). **b**, Electropherograms of *M. sexta sytl* sites L (left) and D (right). Sequences were generated from genomic DNA PCR products (top), RT-PCR products from *M. sexta* brain RNA (middle) and RT-PCR products from *M. sexta* minigene expressed in *Drosophila* S2 cells with dADAR (bottom). **c**, Mutations LM1 and LM2 (as in **a**) were introduced into the *M. sexta* minigene and expressed in S2 cells with dADAR. Values are means $\pm$ s.d.

editing in each domain of the *dsytI* structure can be disrupted independently, it seems unlikely that *dsytI* editing proceeds through a stepwise mechanism invoking sequential editing. Rather, both duplex regions probably exist in the pre-mRNA, forming a long-range pseudoknot containing two domains of ADAR action.

Lepidopterans modify *sytI* at site D, as well as a Lepidoptera-specific site (L). Genomic DNA sequences spanning the same region studied in *Drosophila* species were obtained for ten species of moths and butterflies. Comparison revealed highly conserved intronic sequences downstream of the editing sites, comprising a single extended region with limited sequence variation (Supplementary Fig. S2). To test whether these sequences direct editing, *M. sexta sytI* genomic sequences encompassing the edited exon, downstream intron and downstream exon were expressed as a minigene in *Drosophila* S2 cells along with dADAR. Efficient and specific editing was observed at *M. sexta* sites D and L (Fig. 3b). The *M. sexta* intron is clearly sufficient to direct species-specific editing with a heterologous editing enzyme. Thus, little of the species-specificity of *sytI* editing must be due to differences in ADAR enzymes between species.

Structural predictions for the *M. sexta sytI* pre-mRNA consistently paired conserved intronic sequences with the region of editing sites D and L (Fig. 3a). Although the intronic conserved elements varied in sequence between moth and butterfly species, none of the differences altered the predicted secondary structure of the RNA (Fig. 3, Supplementary Fig. S3). Like *dsytI*, the lepidopteran

substrate contains two duplex domains, each with a site of adenosine modification. Because dADAR efficiently and accurately edits the *M. sexta* minigene in *Drosophila* cells, the predicted structure was tested by mutation (Fig. 3a, c). Two mutations were engineered into intronic conserved sequences (LM1 and LM2). The LM1 mutation decreased editing at site L but editing at site D was relatively unaffected. Mutation LM2 disrupted editing at nearby site D but editing at site L occurred at nearly wild-type levels (Fig. 3c). Thus, *sytI* RNA editing in Lepidoptera occurs through two mutationally separable domains of ADAR action whose overall structural arrangement is substantially different from that seen in *Drosophilidae*.

Anopheles gambiae and *Aedes aegypti sytI* share editing sites with *Drosophila* species (C and D), whereas mosquito-specific site N occurs between sites C and D (Fig. 1c). Because the editing of *Drosophila sytI* at sites C and D is directed by a conserved pseudoknot structure, a comparative sequence analysis of five mosquito species was performed to identify conserved intronic sequences (Supplementary Fig. S4). Like *Drosophila*, mosquitoes possess distinct E1-like and E2-like elements that are predicted to form a homologous pseudoknot structure, with some alterations. Many of these differences are structurally silent; however, key changes are observed in mosquito domain I in the vicinity of editing site C, including the extensive disruption of base pairing near editing site B, which is not edited in mosquitoes (Fig. 4a and Supplementary Fig. S5). Expression of the *Anopheles gambiae sytI*

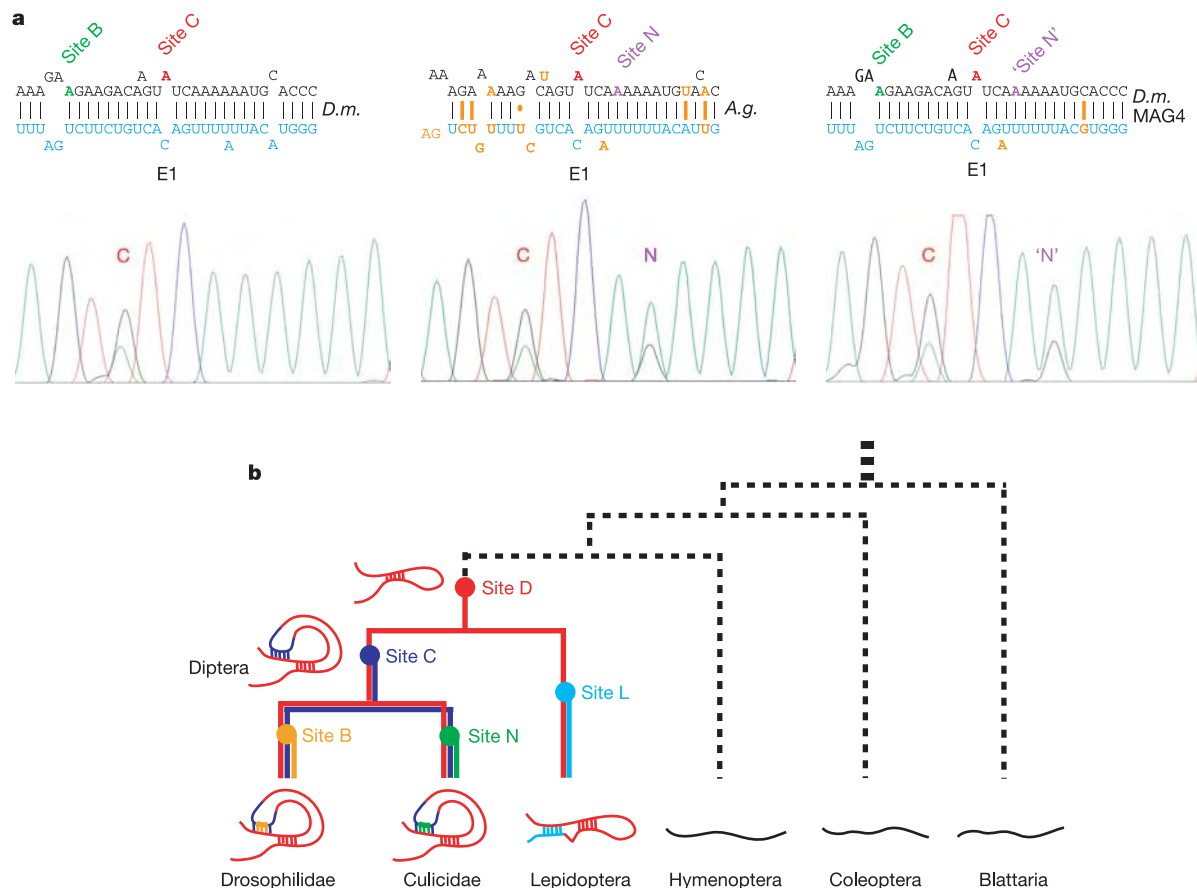


Figure 4 Guided evolution of species-specific RNA editing. **a**, Structure of pseudoknot domain I showing exon (black) and intron (blue) sequences for *Drosophila* (*D.m.*, left), *Anopheles* (*A.g.*, middle) and MAG4 mutation (*D.m. MAG4*, right). Differences of *Anopheles* domain I from *Drosophila* are indicated in orange (middle). The MAG4 mutations (right) are also indicated in orange. Editing status is shown in

electropherograms below structures. **b**, Phylogeny of *sytI* RNA editing. Extant *synaptotagmin I* pre-mRNA structures and proposed ancestor molecules are shown associated with a cladogram of ordinal relationships of taxa in this study³⁰. Nodes denoting ancestral origins of particular editing events are indicated by circles. Unedited *sytI* mRNAs are depicted as unstructured (black).

substrate in *Drosophila* cells along with dADAR resulted in the *Anopheles* pattern of editing, indicating that dADAR recognized the mosquito-specific editing site N (data not shown). To determine whether the editing of site N results from structural differences in domain I, two nucleotide changes were made in the *Drosophila sytI* expression construct in element E1, guided by the *A. gambiae* domain I structure (MAG4; Fig. 4a). The intronic MAG4 mutations are located more than 1,200 nucleotides from editing site C. Whereas no editing of site N occurs in the wild-type *dysytI* construct, the MAG4 mutations confer efficient modification of the mosquito site N adenosine on the *Drosophila* minigene RNA. Thus, small non-coding changes, far removed from a potential target adenosine, are capable of inducing sufficient structural change to directly evolve a site of gene recoding by ADAR modification.

The editing sites described here recode invariant or nearly invariant positions, a phenomenon seen in other targets of A-to-I editing in arthropods, molluscs and vertebrates. Together, these data imply a selective advantage for RNA editing by allowing protein sequences access to a mutational forbidden zone wherein historically invariant amino acids can be altered by degrees, in mRNA, but not through discrete genetic change in coding sequence. Access to this normally unattainable realm of protein space is mediated through complex RNA secondary structures under intense purifying selection. I suggest that the data presented here comprise a credible phylogeny of RNA editing for a gene, graphically illustrating descent with modification (Fig. 4b). RNA editing in insect *sytI* first seems to have evolved in the common ancestor of dipteran and lepidopteran lineages, beginning with site D. Beetles, roaches and reported vertebrate *sytI* genes are genomically incapable of evolving Ile → Met editing at the same location, for lack of a third-position adenosine in their isoleucine codons (ATT or ATC). The ancestral editing site D structure probably generated new editing sites through two methods. One, global intronic variation, led to the synthesis of entirely new, add-on oligonucleotide domains, such as those that direct editing sites C and L in dipterans and lepidopterans, respectively (Figs 2 and 3). Alternatively, sites wended their way through sequence space, generating nascent editing sites by means of a small number of changes to ancestral structures, as shown by the nature of sites A, B and N and the ability to transfer editing from the mosquito to fly pre-mRNA substrate readily by simple mutation. Of course, species-specific *sytI* RNA editing sites, directed by different structures from those presented here, might exist in other animals.

The evolutionary methods of creating editing sites proposed here have probably shaped most targets of ADAR-mediated recoding. Further challenges posed by this study lie in determining what advantage particular editing sites confer on species, the extent of standing variation of RNA editing, and the role, if any, of RNA recoding in the process of speciation. □

Methods

Multiple sequence alignment and structural predictions

Sequences were aligned with the GCG software package. Protein alignment shown in Fig. 1c was performed with the Pileup program using default settings. Sequences are abbreviated as follows: Mmu (*Mus musculus*), Hro (*Halocynthia roretzi*), Aga (*Anopheles gambiae*), Dme (*Drosophila melanogaster*), Ame (*Apis mellifera*), Mse (*Manduca sexta*), Tca (*Tribolium castaneum*), Dja (*Dugesia japonica*) and Lst (*Lymanaea stagnalis*). For *Drosophila* genomic DNA sequence alignments (Supplementary Fig. S1, Supplementary Table S1), alignments were performed on sequences spanning the editing exon through to element E2 because of large discrepancies in intron size downstream of E2. No significant conserved sequences were obtained in alignments downstream of E2 (data not shown). All DNA sequence alignments were performed with the gap creation and gap extension penalties set to a value of 1.

Structural predictions were performed using the MFOLD program of the Macfarlane Burnet Centre MFOLD server (<http://mfold.burnet.edu.au/>), setting the folding temperature to 25 °C. For *Drosophila* substrates, only the sequences from the editing exon as far as 100 nucleotides downstream of E2 were used.

RNA editing analysis of endogenous synaptotagmins

For flies, mosquitoes and beetles, whole-organism RNAs were isolated with TRI Reagent (Molecular Research Center). For moths, butterflies, bees and roaches, whole-head RNA was isolated. Species-specific primers were used for first-strand cDNA synthesis. PCR was then performed with species-specific *sytI* primers. Amplification products of the correct size were gel-purified and sequenced. The presence or absence of editing was assessed by the presence of mixed A/G signals in the electropherograms and singlet A signal from PCR products from genomic DNA for each species.

Schneider cell RNA-editing system and substrate mutagenesis

RNA-editing reporter minigene constructs were generated by cloning the *sytI* editing exon, downstream intron and downstream exon into the pMT-V5/His vector (Invitrogen). dADAR expression construct was generated by cloning the +1, -3a alternative splice form of dADAR into pMT-V5/His. Schneider S2 cells were transfected with various editing reporter constructs (about 200 ng) with or without 2 µg of dADAR expression construct. Cultures were transfected with DNA using GeneJuice (Novagen) transfection reagent and induced with copper sulphate 5 h after transfection. Cells were harvested 3 days after transfection and total RNA was prepared with TRI Reagent. In all cases the RNA samples were treated with DNase (DNA-free; Ambion) to remove contaminating input DNA.

SytI minigene transcripts were amplified by RT-PCR from S2-cell total mRNA with gene-specific and vector-specific primers. RNA editing was quantified as follows. RNA editing of *dysytI* creates a *PstI* restriction enzyme cutting site at site C or a *HpyCH4V* cutting site at site D. *M. sexta* RNA editing creates a *BglI* cleavage site at editing site L and a *HpyCH4V* cutting site at site D. RNA samples were prepared from two to four independent transfections for each construct. For each RNA sample quantified, three independent RT-PCRs were performed with different primer sets; the resulting products were digested with either *PstI*, *BglI* or *HpyCH4V* and subjected to electrophoresis on an agarose gel. The intensities of bands corresponding to edited and unedited products were quantified; band intensities were corrected for band size. Editing frequencies reported are means ± s.d. All images were obtained on a Kodak Gel Logic 100 system and were taken under subsaturation conditions. Data were quantified with 1D v.3.6 image analysis software (Kodak).

Mutations indicated in Figs 2–4 were introduced with PAGE-purified mutagenic primers (about 50 nucleotides in length) (IDT) using the QuikChange II XL Site-directed Mutagenesis Kit (Stratagene). Mutations were induced on the full-length editing constructs in pMT-V5/His. All mutagenized templates were subject to 12–14 rounds of amplification, and the resultant transformed mutants were subjected to sequence analysis of the entire insert to confirm the mutations and lack of secondary mutations.

Received 25 October 2004; accepted 14 January 2005; doi:10.1038/nature03364.

1. Bass, B. L. (ed.) *RNA Editing* (Oxford University Press, Oxford, 2001).
2. Seeburg, P. & Hartner, J. Regulation of ion channel/neurotransmitter receptor function by RNA editing. *Curr. Opin. Neurobiol.* **13**, 279–283 (2003).
3. Hoopengardner, B., Bhalla, T., Staber, C. & Reenan, R. Nervous system targets of RNA editing identified by comparative genomics. *Science* **301**, 832–836 (2003).
4. Bhalla, T., Rosenthal, J. J., Holmgren, M. & Reenan, R. Control of human potassium channel inactivation by editing of a small mRNA hairpin. *Nature Struct. Mol. Biol.* **11**, 950–956 (2004).
5. Herb, A., Higuchi, M., Sprengel, R. & Seeburg, P. Q/R site editing in kainate receptor GluR5 and GluR6 pre-mRNAs requires distant intronic sequences. *Proc. Natl Acad. Sci. USA* **93**, 1875–1880 (1996).
6. Dawson, T., Sansam, C. & Emeson, R. Structure and sequence determinants required for the RNA editing of ADAR2 substrates. *J. Biol. Chem.* **279**, 4941–4951 (2004).
7. Higuchi, M. et al. RNA editing of AMPA receptor subunit GluR-B: a base-paired intron-exon structure determines position and efficiency. *Cell* **75**, 1361–1370 (1993).
8. Bass, B. L. RNA editing by adenosine deaminases that act on RNA. *Annu. Rev. Biochem.* **71**, 817–846 (2002).
9. Tonkin, L. A. et al. RNA editing by ADARs is important for normal behavior in *Caenorhabditis elegans*. *EMBO J.* **21**, 6025–6035 (2002).
10. Palladino, M. J., Keegan, L. P., O'Connell, M. A. & Reenan, R. A. A-to-I pre-mRNA editing in *Drosophila* is primarily involved in adult nervous system function and integrity. *Cell* **102**, 437–449 (2000).
11. Higuchi, M. et al. Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. *Nature* **406**, 78–81 (2000).
12. Gurevich, I. et al. Altered editing of serotonin 2C receptor pre-mRNA in the prefrontal cortex of depressed suicide victims. *Neuron* **34**, 349–356 (2002).
13. Kawahara, Y. et al. Glutamate receptors: RNA editing and death of motor neurons. *Nature* **427**, 801 (2004).
14. Brusa, R. et al. Early-onset epilepsy and postnatal lethality associated with an editing-deficient GluR-B allele in mice. *Science* **270**, 1677–1680 (1995).
15. Vissel, B. et al. The role of RNA editing of kainate receptors in synaptic plasticity and seizures. *Neuron* **29**, 217–227 (2001).
16. Aronoff, R., Mellem, J. E., Maricq, A. V., Sprengel, R. & Seeburg, P. H. Neuronal toxicity in *Caenorhabditis elegans* from an editing site mutant in glutamate receptor channels. *J. Neurosci.* **24**, 8135–8140 (2004).
17. Hanrahan, C. J., Palladino, M. J., Ganetzky, B. & Reenan, R. A. RNA editing of the *Drosophila* para Na⁺ channel transcript. Evolutionary conservation and developmental regulation. *Genetics* **155**, 1149–1160 (2000).
18. Grauso, M., Reenan, R. A., Culetto, E. & Sattelle, D. B. Novel putative nicotinic acetylcholine receptor subunit genes, Dα5, Dα6 and Dα7, in *Drosophila melanogaster* identify a new and highly conserved target of adenosine deaminase acting on RNA-mediated A-to-I pre-mRNA editing. *Genetics* **160**, 1519–1533 (2002).
19. Lehmann, K. A. & Bass, B. L. The importance of internal loops within RNA substrates of ADAR1. *J. Mol. Biol.* **291**, 1–13 (1999).
20. Yoshihara, M. & Littleton, J. T. Synaptotagmin I functions as a calcium sensor to synchronize neurotransmitter release. *Neuron* **36**, 897–908 (2002).

21. Craxton, M. Genomic analysis of synaptotagmin genes. *Genomics* **77**, 43–49 (2001).
22. Mackler, J. M. & Reist, N. E. Mutations in the second C2 domain of synaptotagmin disrupt synaptic transmission at *Drosophila* neuromuscular junctions. *J. Comp. Neurol.* **436**, 4–16 (2001).
23. Chapman, E. R., Desai, R. C., Davis, A. F. & Tornehl, C. K. Delineation of the oligomerization, AP-2 binding, and synprint binding region of the C2B domain of synaptotagmin. *J. Biol. Chem.* **273**, 32966–32972 (1998).
24. Rickman, C. *et al.* Synaptotagmin interaction with the syntaxin/SNAP-25 dimer is mediated by an evolutionarily conserved motif and is sensitive to inositol hexakisphosphate. *J. Biol. Chem.* **279**, 12574–12579 (2004).
25. Grass, I., Thiel, S., Honing, S. & Haucke, V. Recognition of a basic AP-2 binding motif within the C2B domain of synaptotagmin is dependent on multimerization. *J. Biol. Chem.* **279**, 54872–54880 (2004).
26. Nakhost, A., Houeland, G., Blandford, V. E., Castellucci, V. F. & Sossin, W. S. Identification and characterization of a novel C2B splice variant of synaptotagmin I. *J. Neurochem.* **89**, 354–363 (2004).
27. Reenan, R., Hanrahan, C. & Ganetzky, B. The mle(napts) RNA helicase mutation in *Drosophila* results in a splicing catastrophe of the *para* Na⁺ channel transcript in a region of RNA editing. *Neuron* **25**, 139–149 (2000).
28. Aruscavage, P. & Bass, B. A phylogenetic analysis reveals an unusual sequence conservation within introns involved in RNA editing. *RNA* **6**, 257–269 (2000).
29. Kung, S. S., Chen, Y. C., Lin, W. H., Chen, C. C. & Chow, W. Y. Q/R RNA editing of the AMPA receptor subunit 2 (GRIA2) transcript evolves no later than the appearance of cartilaginous fishes. *FEBS Lett.* **509**, 277–281 (2001).
30. Wheeler, W. C., Whiting, M., Wheeler, Q. D. & Carpernter, J. M. The phylogeny of the extant hexapod orders. *Cladistics* **17**, 113–169 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements I thank L. Reenan for discussions; B. Hoopengardner, T. Bhalla and A. Das for comments on the manuscript; B. Hoopengardner for sharing certain genomic DNA templates and for assistance with S2 cell culture; UCHC Molecular Core Facility staff for diligent sequencing efforts; and M. Lalande for his encouragement. This work was supported by grants from the National Science Foundation and National Institutes of Health (R.A.R.).

Competing interests statement The author declares that he has no competing financial interests.

Correspondence and requests for materials should be addressed to R.A.R. (reenan@neuron.uchc.edu).

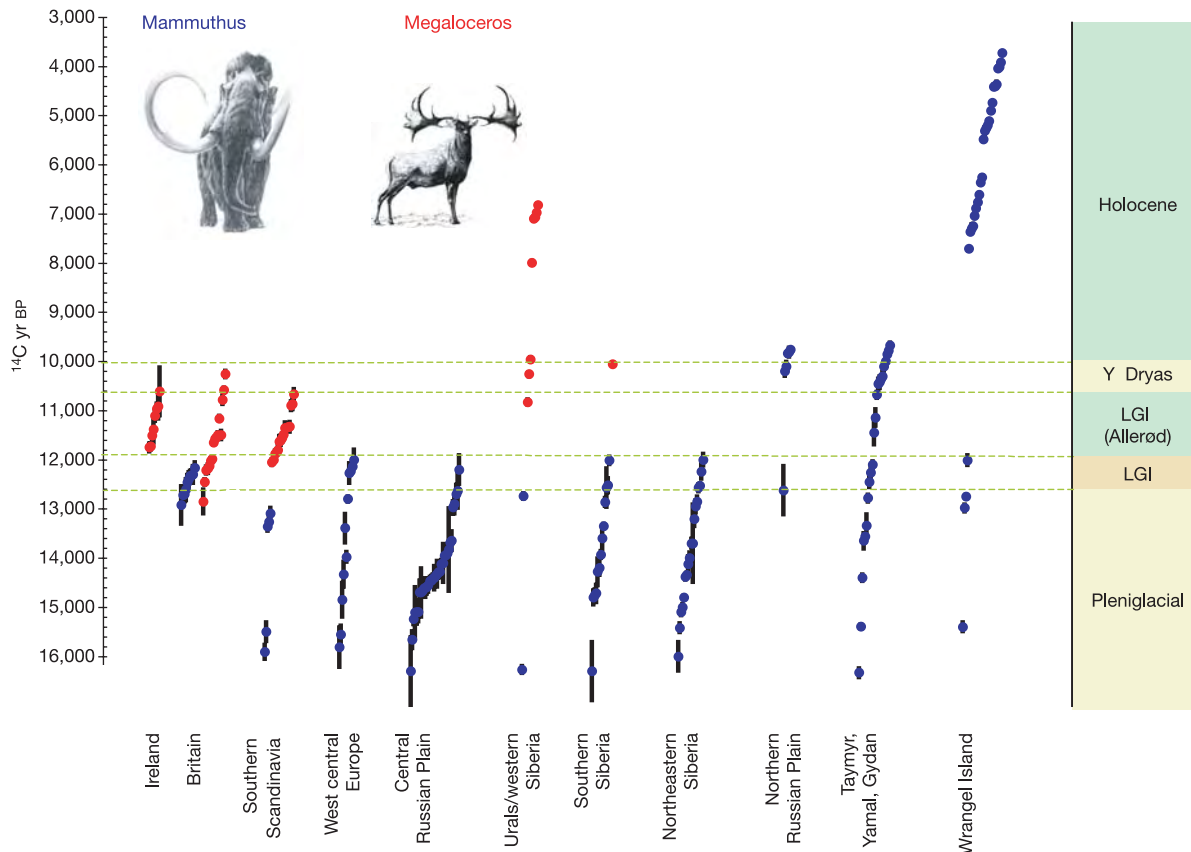
erratum

Pleistocene to Holocene extinction dynamics in giant deer and woolly mammoth

A. J. Stuart, P. A. Kosintsev, T. F. G. Higham & A. M. Lister

Nature **431**, 684–689 (2004).

In Fig. 4 of this Letter, some of the data were not properly aligned with their location labels. The corrected figure is shown here. ☐



Contacts

Publisher: Ben Crowe
Editor: Paul Smaglik
Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs Sales Director:

Nevin Bayouni (4978)

European Sales Manager:

Andy Douglas (4975)

UK/ RoW/ Ireland:

Nils Moeller (4953)
Irene Vigila-Alton (4944)

Scandinavia/ Spain/ Portugal:

Evelina Rubio Håkansson (4973)

Natureevents:

Sille Opstrup (4994)

France/ Switzerland:

Amelie Pequignot (4974)

Advertising Production

Manager: Billie Franklin

To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Niamh Shields

European Satellite Office

Germany/ Austria/ Italy/

The Netherlands/ Belgium:

Patrick Phelan

Tel +49 89 54 90 57 11

Fax +49 89 54 90 57 20

e-mail: p.phelan@nature.com

Reya Silao

Tel +44 (0) 20 7843 4970

e-mail: rsilao@nature.com

US Head Office, New York

345 Park Avenue South,

10th Floor, New York, NY 10010-1707

Tel +1 800 989 7718

Fax +1 800 989 7103

e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

Japan Head Office, Tokyo

MG Ichigaya Building (5F),

19-1 Harakakamachi,

Shinjuku-ku,

Tokyo 162-0841

Tel +81 3 3267 8751

Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami

e-mail: rasami@naturejpn.com

naturejobs

Postdocs, mentor thyself

I thought I was on pretty firm ground when I addressed postdocs at the Salk Institute in San Diego, California, about what young scientists should expect from their principal investigator (PI). The minimum standard, I said, would be someone who trains postdocs and graduate students in scientific skills such as lab techniques, experimental design and data analysis. A conscientious PI would help postdocs advance their scientific careers by developing their non-scientific skills such as grant-writing and lab management. And an exemplary PI would be sympathetic to their off-the-bench goals — even offering to write a letter of reference for them if they are seeking a non-scientific industrial job. Some members of the postdoc association present chuckled at the first, laughed at the second and were practically in hysterics by the third. That would have been edifying if I had intended the round-table discussion of postdoc issues and concerns to be a comedy workshop.

The laughter was discouraging, because it signalled what I've heard all too often at similar settings — that postdocs are considered by many PIs to be just 'an extra pair of hands'. But many of the fellows at the meeting had at least a partial solution to their PIs being less than hands-on in terms of training. They offer to help train their postdoc peers.

Having a formal postdoc organization can help fellows gain non-scientific skills and key career tips. For example, one fellow at the round table was wondering how to get started in regulatory affairs, and a colleague suggested a fellowship with the US Food and Drug Administration. *Naturejobs* has also offered a look into alternative career pathways. And, starting this issue, we are launching a monthly series on how to acquire important non-scientific skills, beginning with presentations (see page 416). This series aims to provide ways to let postdocs help each other and themselves gain these skills — because being unprepared is no laughing matter.

Paul Smaglik
Naturejobs editor



Contents

POSTDOCS AND STUDENTS

The fine art of giving a talk p416

CAREER VIEW

Scientists and Societies

Community outreach

Graduate Journal

Writer's block

Movers

Barry Dickson p418

WWW.NATUREJOBS.COM

Career centre

Information on the scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

Make your point

Engaging presentations can land jobs and increase visibility. Kendall Powell learns the art of a good talk.



A nightmare unfolded as Shawn Hayes presented his graduate work at the annual meeting of the Federation of American Societies for Experimental Biology. His computer crashed just as he was starting his 12-minute talk, requiring a 4-minute reboot. With one hand Hayes tried to find what was wrong with his laptop, while with the other he gestured to the audience as he explained figures from his first slides.

"I was able to stay on track because I had memorized my talk," says Hayes, now an adjunct professor of physiology at the University of California, Davis. "All the big people in my field were going to be there so I wanted to know it backwards and forwards." The preparation paid off — he recovered smoothly from what could have been a talk-ending glitch, and subsequently received several offers of postdoc posts.

Giving forethought to the structure and delivery of a presentation can make the difference between a memorable display of your work and a dud.

Experienced faculty members say that a talk reflects on the scientist who delivers it, not just his or her science.

Each seminar, job talk or poster presentation is a chance to impress future employers, collaborators and even tenure and grant reviewers. Practical advice for creating the best impression may seem obvious, but it's worth remembering that even small *faux pas* can derail intriguing data. To prevent those embarrassments — and to turn talks into career opportunities — consider who is in your audience, what story you're trying to tell, what devices, such as analogies or examples, you'll need, and what style of delivery you will use. And to make all these elements fall into place, practise.

Tailor the presentation to fit the technical level of those listening. Erica Ollmann Saphire, an assistant professor at the Scripps Research Institute in La Jolla, California, suggests that speakers should put

themselves in the listener's shoes. This will also help them to place key points in the most logical order.

Ollmann Saphire studies protein structures from the Ebola and dengue viruses to try to determine what makes these haemorrhagic fevers so pathogenic. She often talks to mixed audiences of immunologists and X-ray crystallographers. In these circumstances she uses really obvious slides such as "Now to immunology", to mark transitions in the subject matter of her talks.

Even subtle courtesies can help get the audience on your side. David Botstein, director of the Lewis-Sigler Institute for Integrative Genomics at Princeton University in New Jersey, recently delivered a seminar at an American Association for the Advancement of Science meeting in Washington DC on interpreting microarray data. He knew that some of his large audience would be colour blind, so for each microarray slide, where gene expression is usually illustrated in red and green, he provided a duplicate in blue and yellow.

SPINNING A SCIENTIFIC TALE

Once you know your topic and audience, remind yourself that a talk is a narrative — not just a collection of facts and figures. "When people go to a seminar, they want a story," says Robert Anholt, a neuroscientist at North Carolina State University in Raleigh. Just as a good tale usually has one moral, he says, a good talk should have one clear, take-home message. Young researchers often make the mistake of trying to cover several projects in a 15-minute talk, he adds.

Using a quote or tagline sentence conveys the overall point. In one memorable departmental seminar, Anholt recalls, a speaker used a quote to illustrate the importance of regulatory gene sequences: "If you give me the coding sequences of a chimp and the regulatory sequences of a mouse, I'll give you a mouse."

For overall structure, Anholt urges, engage your



Speaking out: (from far left) Robert Anholt, Michael De Robertis and Erica Ollmann Saphire all agree preparation is the key to a successful science talk.

audience with the ‘zoom in, zoom out’ strategy. Start with the broad implications of a scientific problem and move in from there to talk about specific data. At the end, at least briefly, zoom back out. Michael De Robertis, an astronomer at York University in Toronto, Canada, recommends placing data into context, showing how they have advanced the field and suggesting what challenges remain.

Once you’ve found your central message, you can plan elegant ways to drive it home. The best speakers use word pictures that illuminate, rather than duplicate, their slides. Botstein, who says that teaching undergraduates greatly improves speaking skills, used an effective analogy in a recent talk on microarray data patterns. “If you are the Duke of Wellington at Waterloo looking through your spyglass at 10,000 soldiers, are you going to see one soldier making a flanking manoeuvre or are you going to see a whole regiment?” The phrase “regiments of genes” used later in the talk drove home his pattern-finding point.

Judith Campisi used another explanatory strategy during a symposium at the 2004 American Society of Cell Biology meeting in Washington DC to explain the concept of ‘antagonistic pleiotropy’. As the term related to her work on cell senescence and cancer, she explained: “In other words, what’s good for you when you are young might be bad for you when you are old.”

Once you’ve found the right words, remember that it’s not just what you say, it’s how you say it, both in terms of verbal and non-verbal traits. Your delivery can leave lasting impressions, both good and bad. The right amount of enthusiasm conveys that a speaker thinks the science “is worth doing, likes doing it, and would like doing it in your lab or department, too”, De Robertis says. Alternatively, he says, “no matter who you are, if you go over the time limit, the majority of the people in the room start tensing up and erasing the talk from their memory”.

Exceeding the allotted time is the number one cardinal sin of delivering a talk (see ‘Top 5 worst speaker sins’, right). It signals a failure to plan and it tramples on the goodwill of the audience. Another no-no is using a rising intonation in your voice, turning every sentence into a question.

Visual aids should be free of distracting colour

schemes or animation. As you describe them, do not turn towards the screen and away from the microphone. Dress appropriately — it says that you care about your presentation and are honoured to speak. For clip-on microphones, it is best to have a collared shirt and a pocket for the battery pack. Taking care of these details eliminates unnecessary disruptions.

No matter how much anyone attends to the elements of a talk, it is all for naught if you don’t practise. Tom Perkins, a biophysicist at the University of Colorado, Boulder, offers an incentive to take preparation and practice seriously. “All talks are potentially job talks,” he says.

Kendall Powell is a freelance science writer in Broomfield, Colorado.

TOP 5 WORST SPEAKER SINS

1

Exceeding the time limit

“I feel like I’m being held prisoner by a person who couldn’t prioritize.” — Erica Ollmann Saphire, Scripps Research Institute

2

An out of control laser pointer

“The audience’s eyes are tracking the spot whizzing around the screen.” — Tom Perkins, University of Colorado, Boulder

3

The rising intonation? Okay? See what I mean?

“This drives me up the wall.” — Robert Anholt, North Carolina State University

4

Speaking too quietly, too fast or incoherently

“If I can’t hear them it is just a waste of my time and I’m going to walk out.” — Shawn Hayes, University of California, Davis

5

Being ambushed by an overeager poster presenter

“It’s like being a good waiter: a poster presenter should be there when you need them, but not be intrusive.” — Robert Anholt, North Carolina State University

GRADUATE JOURNAL

Writer's block

Whether for the advancement of science, of one's career or of a controversial theory, the importance of publishing remains incontrovertible. Although I'd sometimes like to ignore its significance, I can't.

Especially now, when I find myself facing the task of writing my first paper. The data are nearly there. It's almost the interesting story I imagined it would be when I first began the project two years ago. But it's a challenge to put the story into a package that the rest of the scientific community will find interesting.

To do that, I need to understand its place in the context of what has come before. I need to highlight how my story is new and different — and why it deserves to be published. At the very least, I have to convince unknown peer reviewers that my data add something of importance to the collective knowledge.

Having never written a paper before, I wonder if I will be able to do this convincingly at all. It may be that my anxiety at not being able to do so is the reason I haven't started writing yet. And I wonder whether my increasing anxiety should be excitement instead.

Anxiety and doubts or not, I know that I have to start writing. Otherwise, my work will remain unpublished — a nightmare for any ambitious researcher. ■

Anne Margaret Lee is a graduate student at Harvard University.

SCIENTISTS & SOCIETIES

Community outreach

The general public often sees research universities as isolated entities that lack strong relationships with their surrounding communities. This perception, whether deserved or not, has given the universities the unfortunate image of ivory towers of academic isolation. In some cases, this has led to an underrepresentation of minority students in the sciences. To change this perception and increase the diversity in the sciences, it is increasingly important for scientists to initiate and participate in outreach projects involving the community at large.

At the University of California, Los Angeles, we are trying to address this issue in collaboration with the African American Male Achievers Network (A-MAN). Established by Hal Walker, a laser-systems specialist who collaborated

on NASA's first manned Moon mission, and Bettye Walker, a long-time Los Angeles educator, A-MAN was set up to increase the presence of minority students in the natural and physical sciences.

At Los Angeles, we have focused on teaching local middle- and high-school students what it is like to be a scientist at the university, by inviting them to work with us in the lab. We have introduced the students to the fundamentals of biological research, and have nurtured their natural curiosity through scientific experimentation. This year, we have performed experiments ranging from the analysis of bacterial morphology to DNA isolation and analysis. The students have even learned how to grow protein crystals for use in X-ray crystallography.

It is inspiring to see the students' enthusiasm and natural curiosity blossom during our weekly sessions,

and it is gratifying to hear them exclaim "cool!" when we demonstrate new experiments.

Experimental procedures that seem standard to most working scientists take on a new life when students do them for the first time. We hope that our work with these students will encourage them to continue their pursuit of scientific discovery, and as a result, increase the diversity of the next generation of creative scientists.

Science can only benefit from creating such bonds between the university and the surrounding community. Together, these bonds are likely to form long-lasting relationships that will undoubtedly foster scientific progress. ■

Michael Strong participates in the A-MAN community outreach project at the University of California, Los Angeles.

► www.aman.org

► www.doe-mbi.ucla.edu/AMAN/index.html

MOVERS Barry Dickson, director, Research Institute of Molecular Pathology, Vienna, Austria



One book changed the course of Barry Dickson's career. Soon after the Australian researcher finished reading *Eighth Day of Creation*, Horace Freeland Judson's modern history of molecular biology, he swapped mathematics for biology. "Without it I wouldn't have become a biologist," says Dickson, who last month was appointed director of the Research Institute of Molecular Pathology (IMP) in Vienna.

The book may have sparked Dickson's fascination with the molecular mechanisms in cells, but it was good mentors who helped to convince him

that he had made the right move, as they helped to steer his research interests towards problems that excited him.

Ernst Hafen, a fruitfly geneticist who became his supervisor at the University of Zurich in Switzerland, taught him the joy of practising science. And during a postdoctoral stay with Corey Goodman at the University of California, Berkeley, Dickson finally made contact with neurobiology. This was the time he discovered what he really wanted to do: study developmental genetics and its significance for animal behaviour.

Aged 42, Dickson will take up his position at the IMP next January. Together with the Institute of Molecular Biotechnology, the IMP forms the backbone of Vienna's new Genome Research Center.

Dickson, who has worked at both institutes, will step into the shoes of Kim Nasmyth, a British biochemist who over the past few years has led the IMP to the

forefront of molecular biology. Dickson wants to continue his predecessor's policy of nurturing scientific talent and investing in young, ambitious people with a strong affinity for basic research.

The IMP is a good place to train young scientists, Dickson says, because they don't need to write grants every year to get funded, as the institute is financed by the drug company Boehringer Ingelheim. That freedom allows young scientists to look for really important problems to tackle and not be distracted by a constant pressure to publish.

In his new position, Dickson is unlikely to have much free time to engage in the reading that changed his life when he was a student. But in a sense, the new and the unexpected have always been present in his career, as his latest move proves. "When I first heard that I would become the new director of the IMP, I nearly fell off my chair," he says. ■

CV **2003–05:** Senior scientist, Institute of Molecular Biotechnology, Austrian Academy of Sciences, Vienna, Austria.
1998–2002: Group leader, Research Institute of Molecular Pathology, Vienna, Austria.
1996–98: Junior group leader, Institute of Zoology, University of Zurich, Switzerland.

Last man standing

Whatever happened to 'boy meets girl'?

Xaviera Young

She lay with her head in his lap as the first rays of a rather apocalyptic sunset touched the ocean. Long light slanted over them. He twisted her hair between his fingers and looked rather serious. "Marry me," he said. Gulls cried as they flew over the warm water and the scrubby grass of a brown field. She watched the horizon, shielding her eyes. "Not if you were the last man on Earth," she said, smiling, turning back. She laughed. He laughed. Abruptly they both stopped laughing.

Things had once been good for them both. They had once rolled together in the green grass and giggled madly. She would tease him, he would stare her down with slate-grey eyes that reminded her of storms battering against stony cliffs, and she would give in and give up and all would be bliss. Still, even then she didn't think of marriage. He wasn't ... special enough ... somehow. How things change. In the blink of an eye, she had come to realize how pathetically unspecial she was, really, whereas he had transformed into someone tragically unique. What kind of joke was that?

"I have the virus, you know," he said. She kept her eyes on the Sun. The gulls swooped. She sighed. "I know. You have white spots on the back of your hands. You haven't hidden it very well.

"It's so unfair," she said suddenly, bolting upright, his arm still around her. "Why couldn't you have stayed well just a bit longer? A year. Two. They would have worked out a cure. Don't smirk, they would have. If you'd stuck around to give them a reason to bother. At least now they know what it is."

The Y-Virus. Colloquially known as just 'Y', or 'why' as the various plague-wracked men shrieked on the televised news reports: always good for the ratings. A virus that only hit men. And hit it did. Efficient, ruthless, self-ish, domineering, the Y-virus was; a twist that did not escape those same newsreaders. Jokes were made. Until the scope of it came home. The death toll climbed relentlessly. The virus chewed its way through chromosomes around the globe, bringing mankind to its knees. There were surprisingly few riots, little looting. Conflicts came to an end, high finance geared down a notch. For a while, romances burst into flame as they do in the times before a war: love flickered around the

globe as men and women threw themselves into last-minute passions. Many new lives were started in those months, although many — about half, in fact — never made it. It was sexually transmitted, of course. That was the other joke. Not really very funny.

She flopped back in his lap, and he held her as if she were the one in pain. She cried. But she didn't know for whom. For him? Or for the greater loss. A planet with no more moonlight strolls, not really. No more romantic picnics under bowed, green limbs. No more vicious fights fuelled by lust. "What will happen to us? To all of us?" she sobbed, limp in his arms. "There's only one thing in life, really, one story. Boy meets girl. Always was. What the hell happens to that?"

"It'll be girl meets girl, I suppose. Just as good, maybe."

Men, she despaired. All the same.

Life will go on, of course, she knew. Women slipped into the slots of heads-of-state without too much fuss. And there were clones now, the eldest well into their crucial childbearing years and proving the safety — or at least the viability — of the technology. It would be man's legacy, cloning. Well done, she thought bitterly. Maternity wards had now been handed over to stem-cell research.

Babies were and will be born, many of them much the same as their mothers. But what kind of life would that be?

Their sighs floated over the field. A lone cow perked its ears and shuffled away. Their pain was palpable. He slipped his hand into hers, and brushed a stray strand of hair from her face. "You could come with me you know."

She spun the ring on her finger and stared at the Sun, blood red on the horizon. Pollution made for such lovely sunsets. The world was getting better, in some ways. With a 50% cut in population came instability, but also the end of famine. Carbon emissions were plummeting. Not that any of that made any real difference to her, she had to admit. It never did. Tens of thousands died every day in the good old days — without food, choking in cities — and she never noticed or really, to be honest, much cared. She noticed those dying now. She didn't like it.

He opened his palm to reveal a handful of white pills, as the last gasp of a rather apocalyptic sunset slipped over the horizon.

She swallowed. He swallowed. It was the end of the last and only story. ■

Xaviera Young is a freelance writer based in Vancouver, Canada.



JACEY